2_1_Dataset_Preparation_with_Re_balanced_Split_FINAL_ipynb

April 13, 2025

This notebook completes all of the same tasks as 2.0 but with one primary difference.

We re-balanced the dataset, such that one record from the train sets of both the single and multi tasks were stochastically tested so that the following constraints were maintained:

- 1) the values of each sub-set's quartiles remained within 1 standard deviation of each other (just as in the original dataset's close relationship between continuous quartile values)
- 2) the balance of the counts of binarized values remain similarly close together, as in the original dataset
- 3) the balance of counts of subcorpora remain similarly close, as in the original dataset

Originally the single task had: - 7662 train records - 421 validation records - 917 test records

We re-balanced to have in the single set: - 7000 train records - 1000 validation records - 1000 test records

For the multi set, we originally had: - 1517 train records - 99 validation records - 184 test records

We re-balanced to have in the single set: - 1300 train records - 250 validation records - 250 test records

This additional work was performed in order to rule out dataset imbalance issues (predominantly in the validation set) causing 100% recall and precision scores, and highly consistent test F1 scores, across multiple training runs on multiple distinct model architectures

[ ]:
```
#@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
```

```
                         491.2/491.2 kB
8.2 MB/s eta 0:00:00
                         116.3/116.3 kB
2.2 MB/s eta 0:00:00
                         183.9/183.9 kB
10.6 MB/s eta 0:00:00
                         143.5/143.5 kB
```

```
2.9 MB/s eta 0:00:00
                            194.8/194.8 kB
3.6 MB/s eta 0:00:00
                            289.9/289.9 kB
3.8 MB/s eta 0:00:00
                            118.3/118.3 kB
5.5 MB/s eta 0:00:00
```

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

```
Hit:1 http://archive.ubuntu.com/ubuntu jammy InRelease
Get:2 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [128 kB]
Get:3 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease
[3,632 B]
Get:4 http://security.ubuntu.com/ubuntu jammy-security InRelease [129 kB]
Hit:5 http://archive.ubuntu.com/ubuntu jammy-backports InRelease
Get:6 https://r2u.stat.illinois.edu/ubuntu jammy InRelease [6,555 B]
Get:7 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
[18.1 kB]
Hit:8 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Get:9 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [3,099
kB]
Get:10 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 Packages
[1,542 kB]
Get:11 https://r2u.stat.illinois.edu/ubuntu jammy/main all Packages [8,833 kB]
Get:12 http://security.ubuntu.com/ubuntu jammy-security/main amd64 Packages
[2,788 kB]
Get:13 https://r2u.stat.illinois.edu/ubuntu jammy/main amd64 Packages [2,690 kB]
Get:14 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy/main amd64
Packages [34.3 kB]
Get:15 http://security.ubuntu.com/ubuntu jammy-security/universe amd64 Packages
[1,243 kB]
Fetched 20.5 MB in 3s (7,964 kB/s)
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
The following NEW packages will be installed:
  tree
0 upgraded, 1 newly installed, 0 to remove and 2 not upgraded.
Need to get 47.9 kB of archives.
After this operation, 116 kB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy/universe amd64 tree amd64 2.0.2-1
```

```
[47.9 kB]
Fetched 47.9 kB in 0s (363 kB/s)
debconf: unable to initialize frontend: Dialog
debconf: (No usable dialog-like program is installed, so the dialog based
frontend cannot be used. at /usr/share/perl5/Debconf/FrontEnd/Dialog.pm line 78,
<> line 1.)
debconf: falling back to frontend: Readline
debconf: unable to initialize frontend: Readline
debconf: (This frontend requires a controlling tty.)
debconf: falling back to frontend: Teletype
dpkg-preconfigure: unable to re-open stdin:
Selecting previously unselected package tree.
(Reading database … 122158 files and directories currently installed.)
Preparing to unpack …/tree_2.0.2-1_amd64.deb …
Unpacking tree (2.0.2-1) …
Setting up tree (2.0.2-1) …
Processing triggers for man-db (2.10.2-1) …
```

[ ]:
```
#@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

from collections import defaultdict

import sklearn

import spacy
```

[ ]:
```
# @title Mount Google Drive
```

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

```
Mounted at /content/drive
```

[ ]:
```
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(f"========== {df_name} ==========")

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

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

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

    desc = df['complexity'].describe()  # count, mean, std, min, 25%, 50%, 75%, max
    print("'complexity' Column Stats (incl. quartiles and median):")
    print(desc)

    q1 = desc['25%']
    q2 = desc['50%']  # This is the median
    q3 = desc['75%']
    q_max = desc['max']


    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")

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%: 0.3166666666666666 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
======================================
```

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

```
                                id corpus  \
0  3ZLW647WALVGE8EBR50EGUBPU4P32A  bible
1  34R0BODSP1ZBN3DVY8J8XSIY551E5C  bible
2  3S1WOPCJFGTJU2SGNAN2Y213N6WJE3  bible
3  3BFNCI9LYKQNO9BHXHH9CLSX5KP738  bible
4  3G5RUKN2EC3YIWSKUXZ8ZVH95R49N2  bible

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

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

```
                                id corpus  \
0  3S37Y8CWI8ON8KVM53U4E6JKCDC4WE  bible
1  3WGCNLZJKF877FYC1Q6COKNWTDWD11  bible
2  3UOMW19E6D6WQ5TH2HDD74IVKTP5CB  bible
3  36JW4WBR06KF9AXMUL4N476OMF8FHD  bible
4  3HRWUH63QU2FH9Q8R7MRNFC7JX2N5A  bible

                                          sentence            token  \
0  but the seventh day is a Sabbath to Yahweh you…      seventh day
1  But let each man test his own work, and then h…         own work
2  To him who by understanding made the heavens; …  loving kindness
3  Remember to me, my God, this also, and spare m…  loving kindness
4  Because your loving kindness is better than li…  loving kindness

   complexity
0    0.027778
1    0.050000
2    0.050000
3    0.050000
4    0.075000
```

[ ]:
```
#@title Data Engineering
```

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

0.1 standardize column headers: convert trial_val header from 'subcorpus' 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'})

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]

reference_headers = list(dataframes[0].columns)

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

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

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

0.1.1 Identify if any duplicates exist between sets

[ ]:
```
single_dataframes = [train_single_df, trial_val_single_df, test_single_df]
multi_dataframes = [train_multi_df, trial_val_multi_df, test_multi_df]
single_names = ["train_single_df", "trial_val_single_df", "test_single_df"]
multi_names = ["train_multi_df", "trial_val_multi_df", "test_multi_df"]
for d in single_dataframes:
    d["is_duplicated"] = [{} for _ in range(len(d))]
for d in multi_dataframes:
    d["is_duplicated"] = [{} for _ in range(len(d))]
id_sets_single = {}
for n, d in zip(single_names, single_dataframes):
    id_sets_single[n] = set(d["id"].astype(str).dropna())
id_sets_multi = {}
for n, d in zip(multi_names, multi_dataframes):
    id_sets_multi[n] = set(d["id"].astype(str).dropna())
for df_name, df in zip(single_names, single_dataframes):
    print("Processing", df_name)
    for i in range(len(df)):
        row_id = str(df.loc[i, "id"])
        for other_name, other_df in zip(single_names, single_dataframes):
            if other_name != df_name:
                if row_id in id_sets_single[other_name]:
```

```
                    df.at[i, "is_duplicated"].setdefault(other_name, []).
 ↪append(row_id)
    print("Done", df_name)
for df_name, df in zip(multi_names, multi_dataframes):
    print("Processing", df_name)
    for i in range(len(df)):
        row_id = str(df.loc[i, "id"])
        for other_name, other_df in zip(multi_names, multi_dataframes):
            if other_name != df_name:
                if row_id in id_sets_multi[other_name]:
                    df.at[i, "is_duplicated"].setdefault(other_name, []).
 ↪append(row_id)
    print("Done", df_name)
duplicates_info_single = {}
for df_name, df in zip(single_names, single_dataframes):
    duplicates_info_single[df_name] = sum(len(x)>0 for x in df["is_duplicated"])
duplicates_info_multi = {}
for df_name, df in zip(multi_names, multi_dataframes):
    duplicates_info_multi[df_name] = sum(len(x)>0 for x in df["is_duplicated"])
print("Summary of single df duplicates:", duplicates_info_single)
print("Summary of multi df duplicates:", duplicates_info_multi)
frames_dup = []
for df in single_dataframes:
    frames_dup.append(df[df["is_duplicated"].apply(lambda x: len(x)>0)])
for df in multi_dataframes:
    frames_dup.append(df[df["is_duplicated"].apply(lambda x: len(x)>0)])
duplicate_qa_results_df = pd.concat(frames_dup, ignore_index=True)
print("Duplicates consolidated into duplicate_qa_results_df with shape:",␣
 ↪duplicate_qa_results_df.shape)
```

```
Processing train_single_df
Done train_single_df
Processing trial_val_single_df
Done trial_val_single_df
Processing test_single_df
Done test_single_df
Processing train_multi_df
Done train_multi_df
Processing trial_val_multi_df
Done trial_val_multi_df
Processing test_multi_df
Done test_multi_df
Summary of single df duplicates: {'train_single_df': 0, 'trial_val_single_df':
0, 'test_single_df': 0}
Summary of multi df duplicates: {'train_multi_df': 0, 'trial_val_multi_df': 0,
'test_multi_df': 0}
Duplicates consolidated into duplicate_qa_results_df with shape: (0, 6)
```

- no duplictes exist

[ ]:
```
1000-917
```

[ ]: 83

[ ]:
```
1000-421
```

[ ]: 579

[ ]:
```
83+579
```

[ ]: 662

[ ]:
```
7662-7662
```

[ ]: 0

[ ]:
```
dataframes = [
    ("train_single_df", train_single_df),
    ("train_multi_df", train_multi_df),
    ("trial_val_multi_df", trial_val_multi_df),
    ("test_single_df", test_single_df),
    ("test_multi_df", test_multi_df),
]

expected_corpora = ["bible", "europarl", "biomed"]

overall_sums = defaultdict(int)

for name, df in dataframes:
    counts = df["corpus"].value_counts()

    corpus_counts = counts.reindex(expected_corpora, fill_value=0)

    for c in expected_corpora:
        overall_sums[c] += corpus_counts[c]

    print(f"Counts for {name}:")
    print(corpus_counts)
    print("-" * 40)

print("Overall sums across all dataframes:")
for c in expected_corpora:
    print(f"{c}: {overall_sums[c]}")
```

```
Counts for train_single_df:
corpus
bible       2574
```

```
europarl    2512
biomed      2576
Name: count, dtype: int64
----------------------------------------
Counts for train_multi_df:
corpus
bible       505
europarl    498
biomed      514
Name: count, dtype: int64
----------------------------------------
Counts for trial_val_single_df:
corpus
bible       143
europarl    143
biomed      135
Name: count, dtype: int64
----------------------------------------
Counts for trial_val_multi_df:
corpus
bible       29
europarl    37
biomed      33
Name: count, dtype: int64
----------------------------------------
Counts for test_single_df:
corpus
bible       283
europarl    345
biomed      289
Name: count, dtype: int64
----------------------------------------
Counts for test_multi_df:
corpus
bible       66
europarl    65
biomed      53
Name: count, dtype: int64
----------------------------------------
Overall sums across all dataframes:
bible: 3600
europarl: 3600
biomed: 3600
```

[ ]:
```python
import numpy as np
import pandas as pd
```

```
print("Begin rebalancing multi dataset")
m_needed_test = 250 - len(test_multi_df)
m_needed_val = 250 - len(trial_val_multi_df)
print("Required additions to multi test:", m_needed_test)
print("Required additions to multi validation:", m_needed_val)
m_corpus_vals = train_multi_df["corpus"].unique()
m_q = np.quantile(train_multi_df["complexity"], [0.25, 0.5, 0.75])
print("Multi train quartiles:", m_q)
m_stddev_quart = np.std([
    np.sum((train_multi_df["complexity"] <= m_q[0])),
    np.sum((train_multi_df["complexity"] > m_q[0]) &␣
 ↪(train_multi_df["complexity"] <= m_q[1])),
    np.sum((train_multi_df["complexity"] > m_q[1]) &␣
 ↪(train_multi_df["complexity"] <= m_q[2])),
    np.sum(train_multi_df["complexity"] > m_q[2])
])
print("Multi train quartile count std dev:", m_stddev_quart)
m_test_split = {}
m_val_split = {}
for c in m_corpus_vals:
    m_count_test = int(np.floor(m_needed_test / len(m_corpus_vals)))
    m_count_val = int(np.floor(m_needed_val / len(m_corpus_vals)))
    if c == m_corpus_vals[-1]:
        m_count_test = m_needed_test - sum(m_test_split.values())
        m_count_val = m_needed_val - sum(m_val_split.values())
    m_subset_c = train_multi_df[train_multi_df["corpus"] == c]
    attempts_test = m_subset_c.sample(n=m_count_test, random_state=42,␣
 ↪replace=False) if m_count_test > 0 else m_subset_c.iloc[0:0]
    attempts_val = m_subset_c.drop(attempts_test.index).sample(n=m_count_val,␣
 ↪random_state=84, replace=False) if m_count_val > 0 else m_subset_c.iloc[0:0]
    m_test_split[c] = len(attempts_test)
    m_val_split[c] = len(attempts_val)
    train_multi_df.drop(attempts_test.index, inplace=True)
    train_multi_df.drop(attempts_val.index, inplace=True)
    test_multi_df = pd.concat([test_multi_df, attempts_test], ignore_index=True)
    trial_val_multi_df = pd.concat([trial_val_multi_df, attempts_val],␣
 ↪ignore_index=True)
print("Multi corpus distribution moved to test:", m_test_split)
print("Multi corpus distribution moved to validation:", m_val_split)
print("Checking quartiles across multi dataframes after move")
m_train_counts_quart = [
    np.sum((train_multi_df["complexity"] <= m_q[0])),
    np.sum((train_multi_df["complexity"] > m_q[0]) &␣
 ↪(train_multi_df["complexity"] <= m_q[1])),
    np.sum((train_multi_df["complexity"] > m_q[1]) &␣
 ↪(train_multi_df["complexity"] <= m_q[2])),
```

```
    np.sum(train_multi_df["complexity"] > m_q[2])
]
m_val_counts_quart = [
    np.sum((trial_val_multi_df["complexity"] <= m_q[0])),
    np.sum((trial_val_multi_df["complexity"] > m_q[0]) &␣
 ↪(trial_val_multi_df["complexity"] <= m_q[1])),
    np.sum((trial_val_multi_df["complexity"] > m_q[1]) &␣
 ↪(trial_val_multi_df["complexity"] <= m_q[2])),
    np.sum(trial_val_multi_df["complexity"] > m_q[2])
]
m_test_counts_quart = [
    np.sum((test_multi_df["complexity"] <= m_q[0])),
    np.sum((test_multi_df["complexity"] > m_q[0]) &␣
 ↪(test_multi_df["complexity"] <= m_q[1])),
    np.sum((test_multi_df["complexity"] > m_q[1]) &␣
 ↪(test_multi_df["complexity"] <= m_q[2])),
    np.sum(test_multi_df["complexity"] > m_q[2])
]
print("Multi train quartile distribution:", m_train_counts_quart)
print("Multi validation quartile distribution:", m_val_counts_quart)
print("Multi test quartile distribution:", m_test_counts_quart)
print("Ensuring quartile counts remain within 1 std of original train␣
 ↪distribution")
m_final_std_train = np.std(m_train_counts_quart)
m_final_std_val = np.std(m_val_counts_quart)
m_final_std_test = np.std(m_test_counts_quart)
print("Multi final train quartile count std dev:", m_final_std_train)
print("Multi final validation quartile count std dev:", m_final_std_val)
print("Multi final test quartile count std dev:", m_final_std_test)
print("Summing corpus counts across the six dataframes")
for c in m_corpus_vals:
    s1 = len(train_single_df[train_single_df["corpus"] == c])
    s2 = len(trial_val_single_df[trial_val_single_df["corpus"] == c])
    s3 = len(test_single_df[test_single_df["corpus"] == c])
    m1 = len(train_multi_df[train_multi_df["corpus"] == c])
    m2 = len(trial_val_multi_df[trial_val_multi_df["corpus"] == c])
    m3 = len(test_multi_df[test_multi_df["corpus"] == c])
    print("Corpus", c, "counts across six dataframes:", [s1, s2, s3, m1, m2,␣
 ↪m3])
print("Shuffling multi dataframes")
train_multi_df = train_multi_df.sample(frac=1, random_state=101).
 ↪reset_index(drop=True)
trial_val_multi_df = trial_val_multi_df.sample(frac=1, random_state=102).
 ↪reset_index(drop=True)
test_multi_df = test_multi_df.sample(frac=1, random_state=103).
 ↪reset_index(drop=True)
```

```
print("Finished rebalancing multi dataset")

print("Begin rebalancing single dataset")
s_needed_test = 1000 - len(test_single_df)
s_needed_val = 1000 - len(trial_val_single_df)
print("Required additions to single test:", s_needed_test)
print("Required additions to single validation:", s_needed_val)
s_corpus_vals = train_single_df["corpus"].unique()
s_q = np.quantile(train_single_df["complexity"], [0.25, 0.5, 0.75])
print("Single train quartiles:", s_q)
s_stddev_quart = np.std([
    np.sum((train_single_df["complexity"] <= s_q[0])),
    np.sum((train_single_df["complexity"] > s_q[0]) &␣
 ↪(train_single_df["complexity"] <= s_q[1])),
    np.sum((train_single_df["complexity"] > s_q[1]) &␣
 ↪(train_single_df["complexity"] <= s_q[2])),
    np.sum(train_single_df["complexity"] > s_q[2])
])
print("Single train quartile count std dev:", s_stddev_quart)
s_test_split = {}
s_val_split = {}
for c in s_corpus_vals:
    s_count_test = int(np.floor(s_needed_test / len(s_corpus_vals)))
    s_count_val = int(np.floor(s_needed_val / len(s_corpus_vals)))
    if c == s_corpus_vals[-1]:
        s_count_test = s_needed_test - sum(s_test_split.values())
        s_count_val = s_needed_val - sum(s_val_split.values())
    s_subset_c = train_single_df[train_single_df["corpus"] == c]
    s_attempts_test = s_subset_c.sample(n=s_count_test, random_state=44,␣
 ↪replace=False) if s_count_test > 0 else s_subset_c.iloc[0:0]
    s_attempts_val = s_subset_c.drop(s_attempts_test.index).
 ↪sample(n=s_count_val, random_state=55, replace=False) if s_count_val > 0␣
 ↪else s_subset_c.iloc[0:0]
    s_test_split[c] = len(s_attempts_test)
    s_val_split[c] = len(s_attempts_val)
    train_single_df.drop(s_attempts_test.index, inplace=True)
    train_single_df.drop(s_attempts_val.index, inplace=True)
    test_single_df = pd.concat([test_single_df, s_attempts_test],␣
 ↪ignore_index=True)
    trial_val_single_df = pd.concat([trial_val_single_df, s_attempts_val],␣
 ↪ignore_index=True)
print("Single corpus distribution moved to test:", s_test_split)
print("Single corpus distribution moved to validation:", s_val_split)
print("Checking quartiles across single dataframes after move")
s_train_counts_quart = [
    np.sum((train_single_df["complexity"] <= s_q[0])),
```

```
    np.sum((train_single_df["complexity"] > s_q[0]) &␣
 ↪(train_single_df["complexity"] <= s_q[1])),
    np.sum((train_single_df["complexity"] > s_q[1]) &␣
 ↪(train_single_df["complexity"] <= s_q[2])),
    np.sum(train_single_df["complexity"] > s_q[2])
]
s_val_counts_quart = [
    np.sum((trial_val_single_df["complexity"] <= s_q[0])),
    np.sum((trial_val_single_df["complexity"] > s_q[0]) &␣
 ↪(trial_val_single_df["complexity"] <= s_q[1])),
    np.sum((trial_val_single_df["complexity"] > s_q[1]) &␣
 ↪(trial_val_single_df["complexity"] <= s_q[2])),
    np.sum(trial_val_single_df["complexity"] > s_q[2])
]
s_test_counts_quart = [
    np.sum((test_single_df["complexity"] <= s_q[0])),
    np.sum((test_single_df["complexity"] > s_q[0]) &␣
 ↪(test_single_df["complexity"] <= s_q[1])),
    np.sum((test_single_df["complexity"] > s_q[1]) &␣
 ↪(test_single_df["complexity"] <= s_q[2])),
    np.sum(test_single_df["complexity"] > s_q[2])
]
print("Single train quartile distribution:", s_train_counts_quart)
print("Single validation quartile distribution:", s_val_counts_quart)
print("Single test quartile distribution:", s_test_counts_quart)
print("Ensuring quartile counts remain within 1 std of original train␣
 ↪distribution")
s_final_std_train = np.std(s_train_counts_quart)
s_final_std_val = np.std(s_val_counts_quart)
s_final_std_test = np.std(s_test_counts_quart)
print("Single final train quartile count std dev:", s_final_std_train)
print("Single final validation quartile count std dev:", s_final_std_val)
print("Single final test quartile count std dev:", s_final_std_test)
print("Summing corpus counts across the six dataframes")
for c in s_corpus_vals:
    s1 = len(train_single_df[train_single_df["corpus"] == c])
    s2 = len(trial_val_single_df[trial_val_single_df["corpus"] == c])
    s3 = len(test_single_df[test_single_df["corpus"] == c])
    m1 = len(train_multi_df[train_multi_df["corpus"] == c])
    m2 = len(trial_val_multi_df[trial_val_multi_df["corpus"] == c])
    m3 = len(test_multi_df[test_multi_df["corpus"] == c])
    print("Corpus", c, "counts across six dataframes:", [s1, s2, s3, m1, m2,␣
 ↪m3])
print("Shuffling single dataframes")
train_single_df = train_single_df.sample(frac=1, random_state=201).
 ↪reset_index(drop=True)
```

```
trial_val_single_df = trial_val_single_df.sample(frac=1, random_state=202).
  ↪reset_index(drop=True)
test_single_df = test_single_df.sample(frac=1, random_state=203).
  ↪reset_index(drop=True)
print("Finished rebalancing single dataset")
```

```
Begin rebalancing multi dataset
Required additions to multi test: 66
Required additions to multi validation: 151
Multi train quartiles: [0.30263158 0.40909091 0.52941176]
Multi train quartile count std dev: 1.920286436967152
Multi corpus distribution moved to test: {'bible': 22, 'biomed': 22, 'europarl':
22}
Multi corpus distribution moved to validation: {'bible': 50, 'biomed': 50,
'europarl': 51}
Checking quartiles across multi dataframes after move
Multi train quartile distribution: [np.int64(324), np.int64(312), np.int64(340),
np.int64(324)]
Multi validation quartile distribution: [np.int64(67), np.int64(71),
np.int64(56), np.int64(56)]
Multi test quartile distribution: [np.int64(56), np.int64(62), np.int64(67),
np.int64(65)]
Ensuring quartile counts remain within 1 std of original train distribution
Multi final train quartile count std dev: 9.9498743710662
Multi final validation quartile count std dev: 6.652067347825035
Multi final test quartile count std dev: 4.153311931459037
Summing corpus counts across the six dataframes
Corpus bible counts across six dataframes: [2574, 143, 283, 433, 79, 88]
Corpus biomed counts across six dataframes: [2576, 135, 289, 442, 83, 75]
Corpus europarl counts across six dataframes: [2512, 143, 345, 425, 88, 87]
Shuffling multi dataframes
Finished rebalancing multi dataset
Begin rebalancing single dataset
Required additions to single test: 83
Required additions to single validation: 579
Single train quartiles: [0.21153846 0.27941176 0.375     ]
Single train quartile count std dev: 62.882827544568954
Single corpus distribution moved to test: {'bible': 27, 'biomed': 27,
'europarl': 29}
Single corpus distribution moved to validation: {'bible': 193, 'biomed': 193,
'europarl': 193}
Checking quartiles across single dataframes after move
Single train quartile distribution: [np.int64(1747), np.int64(1787),
np.int64(1792), np.int64(1674)]
Single validation quartile distribution: [np.int64(256), np.int64(262),
np.int64(270), np.int64(212)]
Single test quartile distribution: [np.int64(252), np.int64(266), np.int64(275),
```

```
np.int64(207)]
Ensuring quartile counts remain within 1 std of original train distribution
Single final train quartile count std dev: 47.217581471312144
Single final validation quartile count std dev: 22.494443758403985
Single final test quartile count std dev: 26.143832924802744
Summing corpus counts across the six dataframes
Corpus bible counts across six dataframes: [2354, 336, 310, 433, 79, 88]
Corpus biomed counts across six dataframes: [2356, 328, 316, 442, 83, 75]
Corpus europarl counts across six dataframes: [2290, 336, 374, 425, 88, 87]
Shuffling single dataframes
Finished rebalancing single dataset
```

[ ]:
```python
dataframes = [
    ("train_single_df", train_single_df),
    ("train_multi_df", train_multi_df),
    ("trial_val_multi_df", trial_val_multi_df),
    ("test_single_df", test_single_df),
    ("test_multi_df", test_multi_df),
]

expected_corpora = ["bible", "europarl", "biomed"]

overall_sums = defaultdict(int)

for name, df in dataframes:
    counts = df["corpus"].value_counts()

    corpus_counts = counts.reindex(expected_corpora, fill_value=0)

    for c in expected_corpora:
        overall_sums[c] += corpus_counts[c]

    print(f"Counts for {name}:")
    print(corpus_counts)
    print("-" * 40)

print("Overall sums across all dataframes:")
for c in expected_corpora:
    print(f"{c}: {overall_sums[c]}")
```

```
Counts for train_single_df:
corpus
bible       2354
europarl    2290
biomed      2356
Name: count, dtype: int64
----------------------------------------
Counts for train_multi_df:
```

```
corpus
bible       433
europarl    425
biomed      442
Name: count, dtype: int64
----------------------------------------
Counts for trial_val_multi_df:
corpus
bible       79
europarl    88
biomed      83
Name: count, dtype: int64
----------------------------------------
Counts for test_single_df:
corpus
bible       310
europarl    374
biomed      316
Name: count, dtype: int64
----------------------------------------
Counts for test_multi_df:
corpus
bible       88
europarl    87
biomed      75
Name: count, dtype: int64
----------------------------------------
Overall sums across all dataframes:
bible: 3264
europarl: 3264
biomed: 3272
```

0.2 Interrogate Span Length by Corpus Value by Data Split

[ ]:
```python
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')
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	33	24.333333	23.0	4.0	48.0	12.516656	125.181818	4.123212
61	test_multi_df	bible	Q2	0.325-0.417	18	18.611111	16.0	6.0	47.0	10.987366	99.666667	4.216215
62	test_multi_df	bible	Q3	0.432-0.528	20	20.350000	21.0	4.0	43.0	10.624079	108.500000	4.423550

```
63        test_multi_df     bible       Q4      0.533-0.694     17  21.411765 ␣
 ↪      20.0        3.0       51.0  12.384763  117.647059      4.556087
64        test_multi_df    biomed       Q1      0.000-0.312     15  26.200000 ␣
 ↪      27.0       10.0       47.0  10.093845  171.000000      5.372335
65        test_multi_df    biomed       Q2      0.324-0.417     13  27.384615 ␣
 ↪      24.0        9.0       47.0  10.484421  174.615385      5.445863
66        test_multi_df    biomed       Q3      0.434-0.528     14  29.714286 ␣
 ↪      26.5       10.0       61.0  13.498881  199.500000      5.624938
67        test_multi_df    biomed       Q4      0.544-0.806     33  31.696970 ␣
 ↪      28.0       14.0       56.0  12.746286  205.181818      5.421726
68        test_multi_df  europarl       Q1      0.172-0.317     17  27.647059 ␣
 ↪      25.0        7.0       59.0  16.066040  165.588235      4.939143
69        test_multi_df  europarl       Q2      0.321-0.422     29  28.103448 ␣
 ↪      28.0        9.0       73.0  14.326162  173.724138      5.279279
70        test_multi_df  europarl       Q3      0.429-0.533     29  32.241379 ␣
 ↪      32.0        6.0       92.0  22.870216  203.034483      5.402966
71        test_multi_df  europarl       Q4      0.536-0.603     12  39.166667 ␣
 ↪      36.5        8.0       95.0  25.171533  237.833333      5.035037
48       test_single_df     bible       Q1      0.000-0.211     85  22.800000 ␣
 ↪      22.0        7.0       49.0  10.346152  116.764706      4.036286
49       test_single_df     bible       Q2      0.212-0.276     79  24.379747 ␣
 ↪      22.0        2.0       77.0  14.174104  125.367089      4.100632
50       test_single_df     bible       Q3      0.278-0.353     71  22.845070 ␣
 ↪      20.0        4.0       63.0  11.438841  122.056338      4.249799
51       test_single_df     bible       Q4      0.359-0.861     75  20.706667 ␣
 ↪      19.0        1.0       55.0  11.294310  111.200000      4.349945
52       test_single_df    biomed       Q1      0.000-0.206     81  27.320988 ␣
 ↪      27.0       10.0       84.0  12.167402  174.320988      5.268403
53       test_single_df    biomed       Q2      0.212-0.275     63  29.666667 ␣
 ↪      26.0       10.0       83.0  15.461711  193.111111      5.410479
54       test_single_df    biomed       Q3      0.278-0.357     73  30.465753 ␣
 ↪      29.0       13.0       85.0  11.761616  195.945205      5.338991
55       test_single_df    biomed       Q4      0.359-0.778     99  30.979798 ␣
 ↪      30.0       14.0       83.0  11.617176  200.474747      5.344383
56       test_single_df  europarl       Q1      0.000-0.211     84  25.464286 ␣
 ↪      22.0        3.0       82.0  15.583778  152.000000      5.014743
57       test_single_df  europarl       Q2      0.212-0.276    111  31.801802 ␣
 ↪      30.0        1.0       97.0  18.329717  192.324324      5.049358
58       test_single_df  europarl       Q3      0.278-0.357    103  32.417476 ␣
 ↪      29.0        3.0      141.0  21.080855  197.766990      5.103406
59       test_single_df  europarl       Q4      0.361-0.583     76  34.828947 ␣
 ↪      29.0        1.0      143.0  23.701133  215.552632      5.161913
12       train_multi_df     bible       Q1      0.028-0.304    142  23.908451 ␣
 ↪      22.0        3.0       67.0  12.460380  126.584507      4.242442
13       train_multi_df     bible       Q2      0.306-0.417    118  23.889831 ␣
 ↪      22.0        5.0       65.0  12.028303  128.567797      4.321466
```

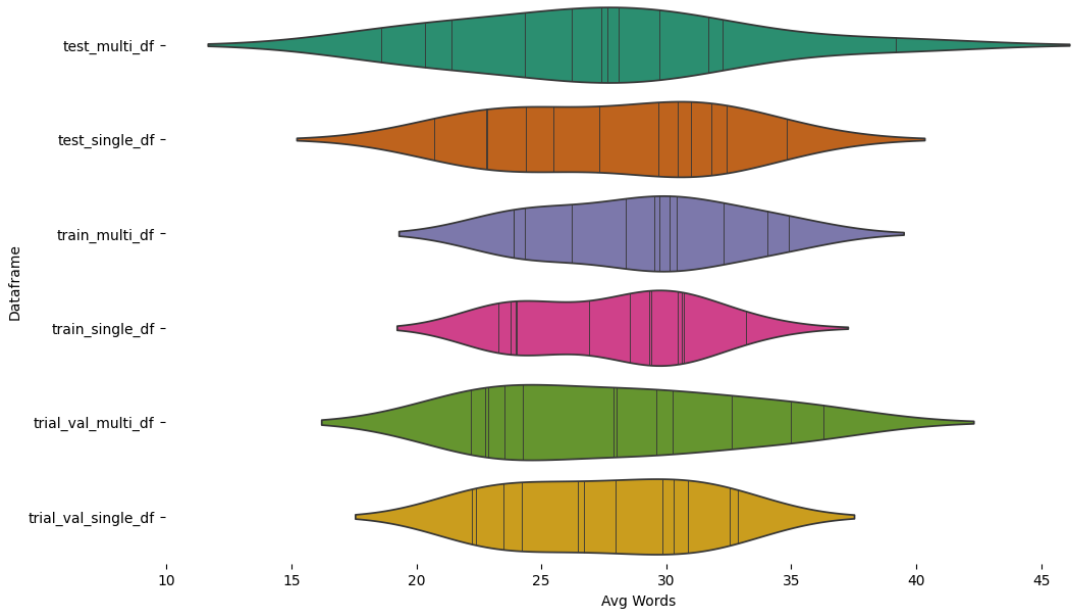
```
14       train_multi_df     bible       Q3      0.420-0.529    107  24.345794 ␣
 ↪      23.0        4.0       50.0  11.186332  130.691589      4.326314
15       train_multi_df     bible       Q4      0.533-0.778     66  26.227273 ␣
 ↪      25.0        4.0       81.0  13.703223  143.833333      4.505703
16       train_multi_df    biomed       Q1      0.028-0.304     74  30.418919 ␣
 ↪      28.5       15.0       77.0  11.645455  195.135135      5.316617
17       train_multi_df    biomed       Q2      0.306-0.417     68  29.720588 ␣
 ↪      27.5       11.0       85.0  13.712230  189.926471      5.410709
18       train_multi_df    biomed       Q3      0.421-0.529     91  30.120879 ␣
 ↪      29.0        8.0       58.0  10.790773  195.956044      5.412654
19       train_multi_df    biomed       Q4      0.531-0.975    209  29.535885 ␣
 ↪      29.0       10.0       75.0  11.946592  194.870813      5.536487
20       train_multi_df  europarl       Q1      0.118-0.304    113  28.389381 ␣
 ↪      24.0        3.0      101.0  18.146445  170.424779      5.019934
21       train_multi_df  europarl       Q2      0.304-0.417    150  32.293333 ␣
 ↪      29.5        3.0       99.0  18.962891  199.300000      5.215569
22       train_multi_df  europarl       Q3      0.420-0.529    113  34.044248 ␣
 ↪      31.0        7.0      101.0  18.924375  210.654867      5.208330
23       train_multi_df  europarl       Q4      0.533-0.750     49  34.897959 ␣
 ↪      31.0        6.0       96.0  21.237982  218.857143      5.308874
0       train_single_df     bible       Q1      0.000-0.212    642  23.302181 ␣
 ↪      22.0        4.0       61.0  11.948497  121.619938      4.125615
1       train_single_df     bible       Q2      0.214-0.279    582  24.003436 ␣
 ↪      22.0        3.0       60.0  11.555150  126.067010      4.147343
2       train_single_df     bible       Q3      0.281-0.375    565  23.971681 ␣
 ↪      22.0        3.0       70.0  12.086681  126.976991      4.204497
3       train_single_df     bible       Q4      0.380-0.825    565  23.778761 ␣
 ↪      21.0        3.0       69.0  12.624089  127.564602      4.289496
4       train_single_df    biomed       Q1      0.000-0.212    537  28.551210 ␣
 ↪      27.0        2.0       85.0  12.249036  181.739292      5.308120
5       train_single_df    biomed       Q2      0.214-0.279    543  30.441989 ␣
 ↪      29.0        7.0       92.0  11.903356  194.040516      5.289866
6       train_single_df    biomed       Q3      0.281-0.375    589  29.407470 ␣
 ↪      28.0        4.0       77.0  11.272763  188.229202      5.329313
7       train_single_df    biomed       Q4      0.381-0.861    687  29.312955 ␣
 ↪      28.0        3.0       85.0  12.433245  187.630277      5.292773
8       train_single_df  europarl       Q1      0.025-0.212    579  26.915371 ␣
 ↪      24.0        2.0      107.0  15.171046  160.526770      4.954404
9       train_single_df  europarl       Q2      0.214-0.279    651  30.626728 ␣
 ↪      28.0        1.0      129.0  18.437599  184.201229      4.981315
10      train_single_df  europarl       Q3      0.281-0.375    638  30.708464 ␣
 ↪      28.0        1.0      122.0  18.318556  187.012539      5.111000
11      train_single_df  europarl       Q4      0.381-0.775    422  33.175355 ␣
 ↪      30.0        2.0      235.0  21.481087  201.936019      5.067161
36  trial_val_multi_df     bible       Q1      0.000-0.292     29  22.758621 ␣
 ↪      20.0        5.0       64.0  12.176251  122.275862      4.229218
```

```
37  trial_val_multi_df     bible       Q2      0.306-0.383     18  22.888889
    21.5        9.0       38.0   8.505285  123.777778      4.288125
38  trial_val_multi_df     bible       Q3      0.389-0.500     17  23.529412
    24.0        5.0       42.0  11.108092  126.823529      4.356087
39  trial_val_multi_df     bible       Q4      0.517-0.661     15  22.200000
    19.0        7.0       44.0  10.448513  118.933333      4.236544
40  trial_val_multi_df    biomed       Q1      0.083-0.297     15  24.266667
    25.0        9.0       49.0  11.460907  149.333333      5.101413
41  trial_val_multi_df    biomed       Q2      0.303-0.383     16  27.875000
    24.5       15.0       54.0  10.843585  174.750000      5.151118
42  trial_val_multi_df    biomed       Q3      0.400-0.513     14  36.285714
    38.0       19.0       48.0   8.686228  229.785714      5.277365
43  trial_val_multi_df    biomed       Q4      0.516-0.825     38  29.605263
    27.5       10.0       70.0  11.785331  197.815789      5.634704
44  trial_val_multi_df  europarl       Q1      0.167-0.298     19  32.631579
    29.0        4.0       67.0  18.083788  196.631579      4.991603
45  trial_val_multi_df  europarl       Q2      0.300-0.383     30  35.000000
    29.0       10.0      108.0  20.857480  219.100000      5.168378
46  trial_val_multi_df  europarl       Q3      0.393-0.515     29  30.241379
    26.0        5.0       78.0  19.279255  194.241379      5.373251
47  trial_val_multi_df  europarl       Q4      0.533-0.714     10  28.000000
    27.0        6.0       66.0  17.606817  185.700000      5.571446
24  trial_val_single_df    bible       Q1      0.000-0.208     98  24.214286
    22.0        5.0       73.0  12.464589  125.846939      4.102514
25  trial_val_single_df    bible       Q2      0.211-0.275     91  23.472527
    22.0        3.0       58.0  12.162182  123.758242      4.233197
26  trial_val_single_df    bible       Q3      0.276-0.359     72  22.388889
    19.5        5.0       45.0  10.253897  119.277778      4.268230
27  trial_val_single_df    bible       Q4      0.361-0.633     75  22.213333
    22.0        4.0       49.0  10.855679  119.213333      4.294968
28  trial_val_single_df   biomed       Q1      0.028-0.206     63  27.968254
    27.0        2.0       65.0  11.689488  181.936508      5.432748
29  trial_val_single_df   biomed       Q2      0.214-0.275     69  30.869565
    30.0       11.0       61.0  11.637810  196.014493      5.245462
30  trial_val_single_df   biomed       Q3      0.278-0.359     90  32.533333
    30.5       10.0       65.0  12.503662  208.422222      5.350775
31  trial_val_single_df   biomed       Q4      0.361-0.875    106  26.716981
    26.0        6.0       77.0  11.900184  174.905660      5.436768
32  trial_val_single_df  europarl       Q1      0.050-0.208     89  26.460674
    24.0        4.0       73.0  16.515183  155.359551      4.894974
33  trial_val_single_df  europarl       Q2      0.211-0.275     93  29.849462
    26.0        4.0      113.0  18.637822  179.451613      5.041307
34  trial_val_single_df  europarl       Q3      0.276-0.359     89  30.314607
    28.0        3.0       99.0  18.045143  184.123596      5.082251
35  trial_val_single_df  europarl       Q4      0.367-0.611     65  32.846154
    31.0        5.0       80.0  18.189309  196.615385      5.002380
```

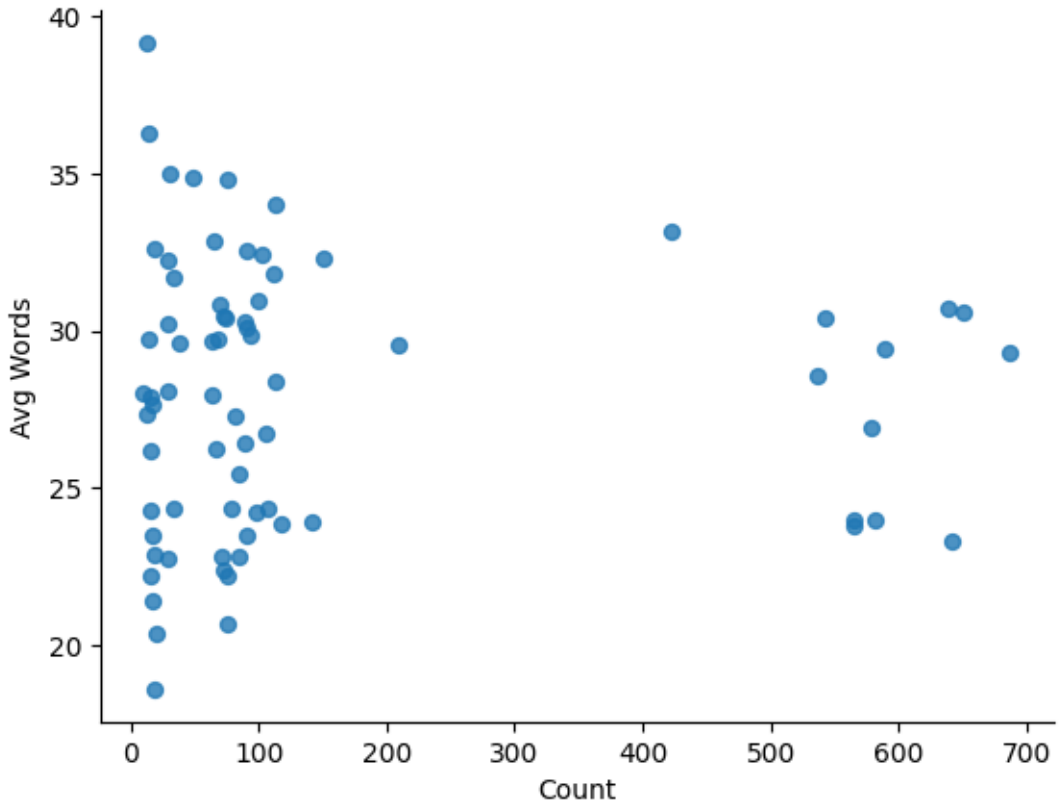
```
Analysis saved to:
/content/drive/MyDrive/266-final/results/sentence_span_analysis.csv
```

[ ]:
```
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-44-00a8ad5642c1>: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')
```

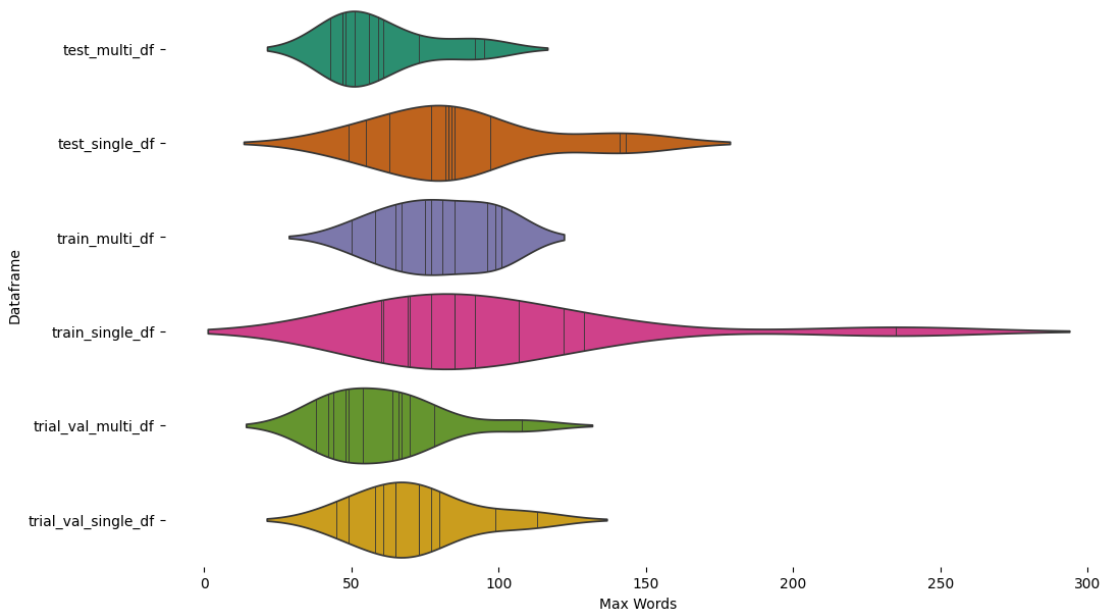


[ ]:
```
from matplotlib import pyplot as plt
span_analysis.plot(kind='scatter', x='Count', y='Avg Words', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```



[ ]:
```
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)
```

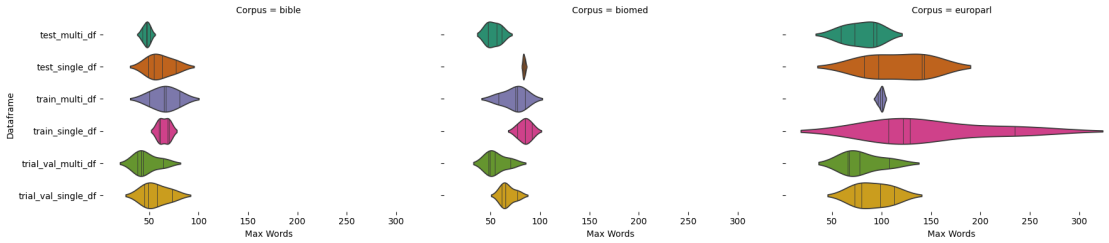
```
<ipython-input-46-01bf0c89d620>:5: FutureWarning:

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')
```



[ ]:
```
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

[ ]:
```python
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 --
       He who is a hired hand, and not a shepherd, who doesn't own the sheep,
sees the wolf coming, leaves the sheep, and flees.
       Bring forth therefore fruits worthy of repentance, and don't begin to say
among yourselves, 'We have Abraham for our father;' for I tell you that God is
able to raise up children to Abraham from these stones!
       But Jonathan didn't hear when his father commanded the people with the
oath: therefore he put forth the end of the rod who was in his hand, and dipped
it in the honeycomb, and put his hand to his mouth; and his eyes were
enlightened.
    -- AFTER  --
       He who is a hired hand, and not a shepherd, who does not own the sheep,
sees the wolf coming, leaves the sheep, and flees.
       Bring forth therefore fruits worthy of repentance, and do not begin to
```

say among yourselves, 'We have Abraham for our father;' for I tell you that God is able to raise up children to Abraham from these stones!

But Jonathan did not hear when his father commanded the people with the oath: therefore he put forth the end of the rod who was in his hand, and dipped it in the honeycomb, and put his hand to his mouth; and his eyes were enlightened.

Corpus: biomed
-- BEFORE --

Epidemiologic assessment of the correlation between a particular variation in DNA sequence, or polymorphism, and risk for BC has been a dominant paradigm for many years.

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).

Through this process, it is also possible that deficits in RanBP2 cause a disturbance in the equilibrium between Cox11, HK1, and RanBP2 by leading to an increase of the inhibitory activity of Cox11 over HKI that promotes the uncoupling of the interaction of HKI from RanBP2, ultimately causing HKI degradation.

-- AFTER --

Epidemiologic assessment of the correlation between a particular variation in DNA sequence, or polymorphism, and risk for BECAUSE has been a dominant paradigm for many years.

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).

Through this process, it is also possible that deficits in RanBP2 because a disturbance in the equilibrium between Cox11, HK1, and RanBP2 by leading to an increase of the inhibitory activity of Cox11 over HKI that promotes the uncoupling of the interaction of HKI from RanBP2, ultimately causing HKI degradation.

Corpus: europarl
-- BEFORE --

(NL) Madam President, ladies and gentlemen, I concur with the rapporteur and the shadow rapporteurs that in order to underline the need for a worldwide agreement on environmental measures within the International Maritime Organization (IMO), a minor amendment of the text is needed.

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

They usually cause problems at work because people do not understand the way they reduce the capacity of sufferers and make them unfit for work.

-- AFTER --

(NL) Madam President, ladies and gentlemen, I concur with the rapporteur and the shadow rapporteurs that in order to underline the need for a worldwide agreement on environmental measures within the International Maritime Organization (I AM GOING TO), a minor amendment of the text is needed.

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

They usually because problems at work because people do not understand the way they reduce the capacity of sufferers and make them unfit for work.

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

Corpus: bible
-- BEFORE --

I hate, I despise your feasts, and I can't stand your solemn assemblies.

Don't turn from it to the right hand or to the left, that you may have good success wherever you go.

Nevertheless these you shall not eat of them that chew the cud, or of those who have the hoof cloven: the camel, and the hare, and the rabbit; because they chew the cud but don't part the hoof, they are unclean to you.

-- AFTER --

I hate, I despise your feasts, and I cannot stand your solemn assemblies.

Do not turn from it to the right hand or to the left, that you may have good success wherever you go.

Nevertheless these you shall not eat of them that chew the cud, or of those who have the hoof cloven: the camel, and the hare, and the rabbit; because they chew the cud but do not part the hoof, they are unclean to you.

Corpus: biomed
-- BEFORE --

The cause of goiter appears to be an impairment of iodide fixation in the follicular lumen due to a reduced rate of iodide transport across the apical membrane of thyroid gland epithelial cells [4].

Furthermore, null mutations in L-Sox5 or Sox-6 cause lethality at or soon after birth, and no effect on cartilage maintenance has been reported (Smits et al. 2001).

Because the FOG2 mutation we report is de novo and the phenotypes of the pulmonary and diaphragmatic defects are similar between mouse and human, we suggest that this mutation in FOG2 is the first reported cause of a human developmental diaphragmatic and pulmonary defect.

-- AFTER --

The because of goiter appears to be an impairment of iodide fixation in

the follicular lumen due to a reduced rate of iodide transport across the apical membrane of thyroid gland epithelial cells [4].

Furthermore, null mutations in L-Sox5 or Sox-6 because lethality at or soon after birth, and no effect on cartilage maintenance has been reported (Smits et al. 2001).

Because the FOG2 mutation we report is de novo and the phenotypes of the pulmonary and diaphragmatic defects are similar between mouse and human, we suggest that this mutation in FOG2 is the first reported because of a human developmental diaphragmatic and pulmonary defect.

Corpus: europarl
-- BEFORE --
However, this unequal trade relationship is not the only cause for concern; another is the case of unsafe products coming from China.

(NL) Madam President, ladies and gentlemen, I concur with the rapporteur and the shadow rapporteurs that in order to underline the need for a worldwide agreement on environmental measures within the International Maritime Organization (IMO), a minor amendment of the text is needed.

(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 --
However, this unequal trade relationship is not the only because for concern; another is the case of unsafe products coming from China.

(NL) Madam President, ladies and gentlemen, I concur with the rapporteur and the shadow rapporteurs that in order to underline the need for a worldwide agreement on environmental measures within the International Maritime Organization (I AM GOING TO), a minor amendment of the text is needed.

(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 --
If the axe is blunt, and one doesn't sharpen the edge, then he must use more strength; but skill brings success.

When they came up out of the water, the Spirit of the Lord caught Philip away, and the eunuch didn't see him any more, for he went on his way rejoicing.

When his speech is charming, don't believe him; for there are seven abominations in his heart.

-- AFTER --
If the axe is blunt, and one does not sharpen the edge, then he must use more strength; but skill brings success.

When they came up out of the water, the Spirit of the Lord caught Philip

away, and the eunuch did not see him any more, for he went on his way rejoicing.
       When his speech is charming, do not believe him; for there are seven abominations in his heart.

  Corpus: biomed
    -- BEFORE --
       TF and EM generated, characterized and maintained the Crx-/- mouse line.
       Here we describe this effort and the discovery of deletion at the ITPR1 locus as a cause of this disorder in mice and of spinocerebellar ataxia 15 (SCA15) in humans.
       Heterozygous mutations of the human PAX6 gene cause aniridia (absence of the iris) and a range of other congenital eye malformations [2].
    -- AFTER --
       TF and THEM generated, characterized and maintained the Crx-/- mouse line.
       Here we describe this effort and the discovery of deletion at the ITPR1 locus as a because of this disorder in mice and of spinocerebellar ataxia 15 (SCA15) in humans.
       Heterozygous mutations of the human PAX6 gene because aniridia (absence of the iris) and a range of other congenital eye malformations [2].

  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.
       It means that, at the very least, we are ensuring that all vessels are insured by solvent insurance companies for the damage they cause, at least within the framework of the IMO conventions.
       The Special Court for Sierra Leone is making a significant contribution to the cause of peace and justice in the Mano River region of West Africa.
    -- 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.
       It means that, at the very least, we are ensuring that all vessels are insured by solvent insurance companies for the damage they because, at least within the framework of the I AM GOING TO conventions.
       The Special Court for Sierra Leone is making a significant contribution to the because of peace and justice in the Mano River region of West Africa.

==
DataFrame: trial_val_multi_df
==

  Corpus: bible

```
    -- BEFORE --
       but I tell you that whoever puts away his wife, except for the cause of
sexual immorality, makes her an adulteress; and whoever marries her when she is
put away commits adultery.
       But when you do merciful deeds, don't let your left hand know what your
right hand does,
       kill utterly the old man, the young man and the virgin, and little
children and women; but don't come near any man on whom is the mark: and begin
at my sanctuary.
    -- AFTER  --
       but I tell you that whoever puts away his wife, except for the because of
sexual immorality, makes her an adulteress; and whoever marries her when she is
put away commits adultery.
       But when you do merciful deeds, do not let your left hand know what your
right hand does,
       kill utterly the old man, the young man and the virgin, and little
children and women; but do not come near any man on whom is the mark: and begin
at my sanctuary.

  Corpus: biomed
    -- BEFORE --
       The cause of goiter appears to be an impairment of iodide fixation in the
follicular lumen due to a reduced rate of iodide transport across the apical
membrane of thyroid gland epithelial cells [4].
    -- AFTER  --
       The because of goiter appears to be an impairment of iodide fixation in
the follicular lumen due to a reduced rate of iodide transport across the apical
membrane of thyroid gland epithelial cells [4].

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

  Corpus: bible
    -- BEFORE --
       Jephthah said to them, "I and my people were at great strife with the
children of Ammon; and when I called you, you didn't save me out of their hand.
       Don't damage the oil and the wine!"
       I, Daniel, alone saw the vision; for the men who were with me didn't see
the vision; but a great quaking fell on them, and they fled to hide themselves.
    -- AFTER  --
       Jephthah said to them, "I and my people were at great strife with the
children of Ammon; and when I called you, you did not save me out of their hand.
       Do not damage the oil and the wine!"
       I, Daniel, alone saw the vision; for the men who were with me did not see
the vision; but a great quaking fell on them, and they fled to hide themselves.

  Corpus: biomed
```

-- BEFORE --

It is expected that the greatest incidence of BC will be among the heaviest smokers.

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.

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 --

It is expected that the greatest incidence of BECAUSE will be among the heaviest smokers.

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.

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 Ţicãyou, 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 --
       Don't count your handmaid for a wicked woman; for I have been speaking
out of the abundance of my complaint and my provocation."
       says Yahweh 'Won't you tremble at my presence, who have placed the sand
for the bound of the sea, by a perpetual decree, that it can't pass it?
       I gave you a land whereon you had not labored, and cities which you
didn't build, and you live in them.
    -- AFTER  --
       Do not count your handmaid for a wicked woman; for I have been speaking
out of the abundance of my complaint and my provocation."
       says Yahweh 'Will not you tremble at my presence, who have placed the
sand for the bound of the sea, by a perpetual decree, that it cannot pass it?
       I gave you a land whereon you had not labored, and cities which you did
not build, and you live in them.

  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.
    -- 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.
```

[ ]:
```
# 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: 230 True values in 'contraction_expanded'
train_multi_df: 44 True values in 'contraction_expanded'
trial_val_single_df: 38 True values in 'contraction_expanded'
trial_val_multi_df: 8 True values in 'contraction_expanded'
test_single_df: 33 True values in 'contraction_expanded'
test_multi_df: 9 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: 7000 entries, 0 to 6999
Data columns (total 8 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   id                         7000 non-null   object
 1   corpus                     7000 non-null   object
 2   sentence                   7000 non-null   object
 3   token                      6995 non-null   object
 4   complexity                 7000 non-null   float64
 5   is_duplicated              7000 non-null   object
 6   sentence_no_contractions   7000 non-null   object
 7   contraction_expanded       7000 non-null   bool
dtypes: bool(1), float64(1), object(6)
memory usage: 389.8+ KB
```

```
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1300 entries, 0 to 1299
Data columns (total 8 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   id                         1300 non-null   object
 1   corpus                     1300 non-null   object
 2   sentence                   1300 non-null   object
 3   token                      1300 non-null   object
 4   complexity                 1300 non-null   float64
 5   is_duplicated              1300 non-null   object
 6   sentence_no_contractions   1300 non-null   object
 7   contraction_expanded       1300 non-null   bool
dtypes: bool(1), float64(1), object(6)
memory usage: 72.5+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   id                         1000 non-null   object
 1   corpus                     1000 non-null   object
 2   sentence                   1000 non-null   object
 3   token                      998 non-null    object
 4   complexity                 1000 non-null   float64
 5   is_duplicated              1000 non-null   object
 6   sentence_no_contractions   1000 non-null   object
 7   contraction_expanded       1000 non-null   bool
dtypes: bool(1), float64(1), object(6)
memory usage: 55.8+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 8 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   id                         250 non-null    object
 1   corpus                     250 non-null    object
 2   sentence                   250 non-null    object
 3   token                      250 non-null    object
 4   complexity                 250 non-null    float64
 5   is_duplicated              250 non-null    object
 6   sentence_no_contractions   250 non-null    object
 7   contraction_expanded       250 non-null    bool
dtypes: bool(1), float64(1), object(6)
memory usage: 14.0+ KB
```

```
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   id                         1000 non-null   object
 1   corpus                     1000 non-null   object
 2   sentence                   1000 non-null   object
 3   token                      1000 non-null   object
 4   complexity                 1000 non-null   float64
 5   is_duplicated              1000 non-null   object
 6   sentence_no_contractions   1000 non-null   object
 7   contraction_expanded       1000 non-null   bool
dtypes: bool(1), float64(1), object(6)
memory usage: 55.8+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 8 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   id                         250 non-null    object
 1   corpus                     250 non-null    object
 2   sentence                   250 non-null    object
 3   token                      250 non-null    object
 4   complexity                 250 non-null    float64
 5   is_duplicated              250 non-null    object
 6   sentence_no_contractions   250 non-null    object
 7   contraction_expanded       250 non-null    bool
dtypes: bool(1), float64(1), object(6)
memory usage: 14.0+ 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 is_duplicated
sentence_no_contractions  contraction_expanded
0  3IQ9OOAYW6ZPOAQ7VNRXLNM4D1DITZ    biomed  The development of sexually
dimorphic reproduc…           organs    0.250000            {}  The
development of sexually dimorphic reproduc…                 False
1  3PA41K45VN4U7YG4VFEGPOVYAII7PP    biomed  We find that the majority of the
```

```
olfactory rec…            usage    0.382353           {}  We find that the
majority of the olfactory rec…                 False
2  36818Z1KV3D5JB9F4KTTMCUN6U7A3I     bible  His lord was angry, and delivered
him to the t…       tormentors    0.328947           {}  His lord was angry,
and delivered him to the t…                False
3  3VJ4PFXFJ37PI5MYJ4PU9LKNJ9SUAF  europarl  The Taiwanese Government has
informed the Coun…  representations    0.315789           {}  The Taiwanese
Government has informed the Coun…                 False
4  37AQKJ12TXOFXO6IPZQ1ZUODOJPTTP  europarl  However, I too want to thank
everyone who took…         relation    0.267857           {}  However, I too
want to thank everyone who took…                 False
                              id    corpus
sentence                       token  complexity is_duplicated
sentence_no_contractions  contraction_expanded
0  3T2EL38U0MK9MPNAD5X3JSYWH8BQXO    biomed  CA = chronic arthritis; CIA =
collagen-induced…      rheumatoid arthritis    0.600000           {}  CA =
chronic arthritis; CIA = collagen-induced…                 False
1  388CL5C1RJN1927IGW7LZKB8JDSLHQ  europarl  Appointments to parliamentary
committees (vote…  parliamentary committees    0.328947           {}
Appointments to parliamentary committees (vote…                False
2  3A3KKYU7P3H3CAKSB7U0000KY58MW4    biomed  The HG9 strain represents a major
epistasis-ba…               mouse model    0.350000           {}  The HG9
strain represents a major epistasis-ba…                False
3  3FBEFUUYRK54GUWXNMRRTF67GLFA6U     bible  For there is an annulling of a
foregoing comma…     foregoing commandment    0.638889           {}  For
there is an annulling of a foregoing comma…                False
4  36QZ6V1589DTI18SO4BLULET5D3SU9     bible  Ezra the priest, with certain heads
of fathers…                  first day    0.116667           {}  Ezra the
priest, with certain heads of fathers…                False
                              id  corpus
sentence    token  complexity is_duplicated
sentence_no_contractions  contraction_expanded
0  3ZQA3IO31BRYBCP1RZKSZEZVXRG1OZ  biomed  In addition to colorectal neoplams,
these indi…  pigment    0.350000           {}  In addition to colorectal
neoplams, these indi…                False
1  3Z3R5YC0P3N5EJOHUFLCIQ9CX7PTFJ   bible  The Queen of the South will rise up
in the jud…     ends    0.302632           {}  The Queen of the South will
rise up in the jud…                False
2  3URJ6VVYUPNF3BMKEH3UXC6Y9BQ4OF  biomed  Since the parental strains differ in
susceptib…    class    0.261905           {}  Since the parental strains
differ in susceptib…                False
3  3MVY4USGB6NO9ADS6NM7BIQIBGKSI1   bible  For the judgment is against you; for
you have …    Tabor    0.633333           {}  For the judgment is against
you; for you have …                False
4  30U1YOGZGAW71ZX6E9LWKLA5JD8SDZ   bible  having a great and high wall; having
twelve ga…   tribes    0.175000           {}  having a great and high wall;
having twelve ga…                False
                              id    corpus
```

```
sentence                 token  complexity is_duplicated
sentence_no_contractions  contraction_expanded
0  3D17ECOUOEV9PNWF8100BB1K20731T     bible  But some of the itinerant Jews,
exorcists, too…       itinerant Jews    0.600000            {}  But some of
the itinerant Jews, exorcists, too…                 False
1  3XBXDSS888JYVS7XL0P726Z273BLXJ  europarl  The next item is the report by
Esther de Lange…       EU legislation    0.285714            {}  The next item
is the report by Esther de Lange…                 False
2  3GITHABACYLNIC7L9OKTP89VZONN2N    biomed  Alternatively, the unusual
transcriptional reg…  olfactory receptors    0.725000            {}
Alternatively, the unusual transcriptional reg…                 False
3  31MCUE39BKM6T2MIQKL3IY5Q4Q13G6    biomed  Genetic disruption of the Dhcr7
results in neo…   neonatal lethality    0.547619            {}  Genetic
disruption of the Dhcr7 results in neo…                 False
4  37PGLWGSJT6QLROK1ED5KWZ8UO3IKA     bible  In it you shall not sow, neither
reap that whi…      undressed vines    0.525000            {}  In it you shall
not sow, neither reap that whi…                 False
                              id    corpus
sentence         token  complexity is_duplicated
sentence_no_contractions  contraction_expanded
0  3ZURAPD288N45ZC8SW12CKQH5QPF1R    biomed  We show that in p150CAF-1-depleted
ES cells, w…  perturbation    0.484375            {}  We show that in
p150CAF-1-depleted ES cells, w…                 False
1  36D1BWBEHN1HOUMLXN5TCTKVUXL2M8    biomed  Lung development is a complex
process that inv…       process    0.250000            {}  Lung development is
a complex process that inv…                 False
2  3QX22DUVOOHQXLKNLXP4EYH6RZBVME  europarl  That is why we want to introduce
the role of m…          role    0.050000            {}  That is why we want to
introduce the role of m…                 False
3  3HXCEECSQMT70MEB5X2ITZH9OICZYL  europarl  (CS) I would just like to emphasise
that this …     groupings    0.210526            {}  (CS) I would just like to
emphasise that this …                 False
4  3WGCNLZJKF877FYC1Q6COKNWTFRD10  europarl  I am from a border county myself
and I am a re…       process    0.183333            {}  I am from a border
county myself and I am a re…                 False
                              id    corpus
sentence                  token  complexity is_duplicated
sentence_no_contractions  contraction_expanded
0  3FK4G712NX0D3OGOBZGLFKW5KGISST     bible  He shall put no oil on it, neither
shall he pu…          sin offering    0.450000            {}  He shall put no
oil on it, neither shall he pu…                 False
1  3UQVX1UPFSHKXGFE8IIVEWDIRVCO2P    biomed  During the last few years the
Wnt1-Cre transge…         powerful tool    0.305556            {}  During the
last few years the Wnt1-Cre transge…                 False
2  3T2EL38U0MK9MPNAD5X3JSYWH9XXQJ  europarl  The next item is the report by Mrs
Fajon, on b…      external borders    0.343750            {}  The next item is
the report by Mrs Fajon, on b…                 False
3  37AQKJ12TXOFXO6IPZQ1ZUODOJMTTM    biomed  The pathogenesis and developmental
```

```
relationshi…  pulmonary hypoplasia    0.675000           {}  The pathogenesis
and developmental relationshi…                  False
4  3NZ1E5QA6Z1DG01BOHHIWKCD28P5B4     bible  Moreover I will make a covenant of
peace with …  everlasting covenant    0.444444           {}  Moreover I will
make a covenant of peace with …                  False
```

[ ]:
```
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	33	24.333333	23.0	4.0	48.0	12.516656	125.272727	4.126658
61	test_multi_df	bible	Q2	0.325-0.417	18	18.611111	16.0	6.0	47.0	10.987366	99.722222	4.218529
62	test_multi_df	bible	Q3	0.432-0.528	20	20.300000	21.0	4.0	43.0	10.578528	108.850000	4.445514
63	test_multi_df	bible	Q4	0.533-0.694	17	21.411765	20.0	3.0	51.0	12.384763	117.647059	4.556087
64	test_multi_df	biomed	Q1	0.000-0.312	15	26.200000	27.0	10.0	47.0	10.093845	171.000000	5.372335
65	test_multi_df	biomed	Q2	0.324-0.417	13	27.384615	24.0	9.0	47.0	10.484421	174.615385	5.445863
66	test_multi_df	biomed	Q3	0.434-0.528	14	29.714286	26.5	10.0	61.0	13.498881	199.500000	5.624938
67	test_multi_df	biomed	Q4	0.544-0.806	33	31.696970	28.0	14.0	56.0	12.746286	205.181818	5.421726
68	test_multi_df	europarl	Q1	0.172-0.317	17	27.647059	25.0	7.0	59.0	16.066040	165.705882	4.942084
69	test_multi_df	europarl	Q2	0.321-0.422	29	28.103448	28.0	9.0	73.0	14.326162	173.724138	5.279279
70	test_multi_df	europarl	Q3	0.429-0.533	29	32.241379	32.0	6.0	92.0	22.870216	203.034483	5.402966
71	test_multi_df	europarl	Q4	0.536-0.603	12	39.166667	36.5	8.0	95.0	25.171533	237.833333	5.035037
48	test_single_df	bible	Q1	0.000-0.211	85	22.788235	22.0	7.0	49.0	10.329217	116.870588	4.044699
49	test_single_df	bible	Q2	0.212-0.276	79	24.379747	22.0	2.0	77.0	14.174104	125.531646	4.106277

```
50       test_single_df     bible       Q3      0.278-0.353     71   22.845070
  ↪      20.0        4.0       63.0  11.438841  122.098592      4.251642
51       test_single_df     bible       Q4      0.359-0.861     75   20.706667
  ↪      19.0        1.0       55.0  11.294310  111.280000      4.353680
52       test_single_df    biomed       Q1      0.000-0.206     81   27.320988
  ↪      27.0       10.0       84.0  12.167402  174.407407      5.273342
53       test_single_df    biomed       Q2      0.212-0.275     63   29.666667
  ↪      26.0       10.0       83.0  15.461711  193.587302      5.421724
54       test_single_df    biomed       Q3      0.278-0.357     73   30.465753
  ↪      29.0       13.0       85.0  11.761616  195.945205      5.338991
55       test_single_df    biomed       Q4      0.359-0.778     99   30.979798
  ↪      30.0       14.0       83.0  11.617176  200.494949      5.344977
56       test_single_df  europarl       Q1      0.000-0.211     84   25.464286
  ↪      22.0        3.0       82.0  15.583778  152.000000      5.014743
57       test_single_df  europarl       Q2      0.212-0.276    111   31.801802
  ↪      30.0        1.0       97.0  18.329717  192.324324      5.049358
58       test_single_df  europarl       Q3      0.278-0.357    103   32.417476
  ↪      29.0        3.0      141.0  21.080855  197.766990      5.103406
59       test_single_df  europarl       Q4      0.361-0.583     76   34.828947
  ↪      29.0        1.0      143.0  23.701133  215.605263      5.162318
12      train_multi_df     bible       Q1      0.028-0.304    142   23.887324
  ↪      22.0        3.0       67.0  12.460490  126.619718      4.247826
13      train_multi_df     bible       Q2      0.306-0.417    118   23.889831
  ↪      22.0        5.0       65.0  12.028303  128.610169      4.322962
14      train_multi_df     bible       Q3      0.420-0.529    107   24.355140
  ↪      23.0        4.0       50.0  11.210888  130.869159      4.334425
15      train_multi_df     bible       Q4      0.533-0.778     66   26.227273
  ↪      25.0        4.0       81.0  13.703223  144.000000      4.510452
16      train_multi_df    biomed       Q1      0.028-0.304     74   30.418919
  ↪      28.5       15.0       77.0  11.645455  195.162162      5.317904
17      train_multi_df    biomed       Q2      0.306-0.417     68   29.764706
  ↪      28.0       11.0       85.0  13.704905  190.147059      5.407432
18      train_multi_df    biomed       Q3      0.421-0.529     91   30.120879
  ↪      29.0        8.0       58.0  10.790773  195.978022      5.413411
19      train_multi_df    biomed       Q4      0.531-0.975    209   29.550239
  ↪      29.0       10.0       75.0  11.934461  194.990431      5.537083
20      train_multi_df  europarl       Q1      0.118-0.304    113   28.389381
  ↪      24.0        3.0      101.0  18.146445  170.442478      5.020739
21      train_multi_df  europarl       Q2      0.304-0.417    150   32.306667
  ↪      29.5        3.0       99.0  18.978245  199.353333      5.215137
22      train_multi_df  europarl       Q3      0.420-0.529    113   34.044248
  ↪      31.0        7.0      101.0  18.924375  210.654867      5.208330
23      train_multi_df  europarl       Q4      0.533-0.750     49   34.897959
  ↪      31.0        6.0       96.0  21.237982  218.857143      5.308874
0      train_single_df     bible       Q1      0.000-0.212    642   23.295950
  ↪      22.0        4.0       61.0  11.952178  121.727414      4.134621
```

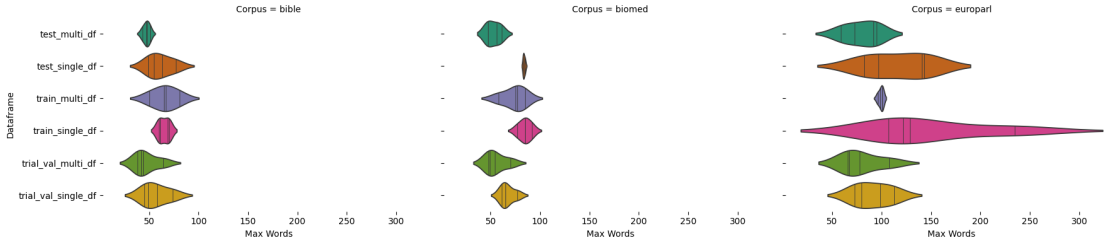
```
1       train_single_df     bible       Q2      0.214-0.279    582  24.000000  ␣
 ↪      22.0        3.0       60.0  11.557087  126.158076      4.152131
2       train_single_df     bible       Q3      0.281-0.375    565  23.973451  ␣
 ↪      22.0        3.0       70.0  12.091451  127.086726      4.210508
3       train_single_df     bible       Q4      0.380-0.825    565  23.785841  ␣
 ↪      21.0        3.0       69.0  12.621824  127.642478      4.291878
4       train_single_df    biomed       Q1      0.000-0.212    537  28.551210  ␣
 ↪      27.0        2.0       85.0  12.249036  181.813780      5.310965
5       train_single_df    biomed       Q2      0.214-0.279    543  30.449355  ␣
 ↪      29.0        7.0       92.0  11.893311  194.174954      5.293255
6       train_single_df    biomed       Q3      0.281-0.375    589  29.407470  ␣
 ↪      28.0        4.0       77.0  11.272763  188.280136      5.331225
7       train_single_df    biomed       Q4      0.381-0.861    687  29.318777  ␣
 ↪      28.0        3.0       85.0  12.436028  187.697234      5.294219
8       train_single_df  europarl       Q1      0.025-0.212    579  26.915371  ␣
 ↪      24.0        2.0      107.0  15.171046  160.537133      4.954812
9       train_single_df  europarl       Q2      0.214-0.279    651  30.626728  ␣
 ↪      28.0        1.0      129.0  18.437599  184.213518      4.981561
10      train_single_df  europarl       Q3      0.281-0.375    638  30.708464  ␣
 ↪      28.0        1.0      122.0  18.318556  187.015674      5.111043
11      train_single_df  europarl       Q4      0.381-0.775    422  33.184834  ␣
 ↪      30.0        2.0      235.0  21.485099  201.981043      5.066969
36   trial_val_multi_df     bible       Q1      0.000-0.292     29  22.758621  ␣
 ↪      20.0        5.0       64.0  12.176251  122.310345      4.230263
37   trial_val_multi_df     bible       Q2      0.306-0.383     18  22.888889  ␣
 ↪      21.5        9.0       38.0   8.505285  124.055556      4.298652
38   trial_val_multi_df     bible       Q3      0.389-0.500     17  23.529412  ␣
 ↪      24.0        5.0       42.0  11.108092  127.058824      4.361825
39   trial_val_multi_df     bible       Q4      0.517-0.661     15  22.200000  ␣
 ↪      19.0        7.0       44.0  10.448513  119.066667      4.242512
40   trial_val_multi_df    biomed       Q1      0.083-0.297     15  24.266667  ␣
 ↪      25.0        9.0       49.0  11.460907  149.333333      5.101413
41   trial_val_multi_df    biomed       Q2      0.303-0.383     16  27.875000  ␣
 ↪      24.5       15.0       54.0  10.843585  174.750000      5.151118
42   trial_val_multi_df    biomed       Q3      0.400-0.513     14  36.285714  ␣
 ↪      38.0       19.0       48.0   8.686228  229.785714      5.277365
43   trial_val_multi_df    biomed       Q4      0.516-0.825     38  29.605263  ␣
 ↪      27.5       10.0       70.0  11.785331  197.868421      5.636252
44   trial_val_multi_df  europarl       Q1      0.167-0.298     19  32.631579  ␣
 ↪      29.0        4.0       67.0  18.083788  196.631579      4.991603
45   trial_val_multi_df  europarl       Q2      0.300-0.383     30  35.000000  ␣
 ↪      29.0       10.0      108.0  20.857480  219.100000      5.168378
46   trial_val_multi_df  europarl       Q3      0.393-0.515     29  30.241379  ␣
 ↪      26.0        5.0       78.0  19.279255  194.241379      5.373251
47   trial_val_multi_df  europarl       Q4      0.533-0.714     10  28.000000  ␣
 ↪      27.0        6.0       66.0  17.606817  185.700000      5.571446
```

```
24  trial_val_single_df     bible       Q1      0.000-0.208     98  24.224490
    ↪      22.0        5.0       74.0  12.505282  126.010204      4.108978
25  trial_val_single_df     bible       Q2      0.211-0.275     91  23.472527
    ↪      22.0        3.0       58.0  12.162182  123.846154      4.238038
26  trial_val_single_df     bible       Q3      0.276-0.359     72  22.388889
    ↪      19.5        5.0       45.0  10.253897  119.375000      4.272391
27  trial_val_single_df     bible       Q4      0.361-0.633     75  22.226667
    ↪      22.0        4.0       49.0  10.862250  119.360000      4.298722
28  trial_val_single_df    biomed       Q1      0.028-0.206     63  27.968254
    ↪      27.0        2.0       65.0  11.689488  181.936508      5.432748
29  trial_val_single_df    biomed       Q2      0.214-0.275     69  30.869565
    ↪      30.0       11.0       61.0  11.637810  196.188406      5.250496
30  trial_val_single_df    biomed       Q3      0.278-0.359     90  32.533333
    ↪      30.5       10.0       65.0  12.503662  208.444444      5.351291
31  trial_val_single_df    biomed       Q4      0.361-0.875    106  26.716981
    ↪      26.0        6.0       77.0  11.900184  174.981132      5.443155
32  trial_val_single_df  europarl       Q1      0.050-0.208     89  26.460674
    ↪      24.0        4.0       73.0  16.515183  155.359551      4.894974
33  trial_val_single_df  europarl       Q2      0.211-0.275     93  29.860215
    ↪      26.0        4.0      113.0  18.650441  179.505376      5.041165
34  trial_val_single_df  europarl       Q3      0.276-0.359     89  30.314607
    ↪      28.0        3.0       99.0  18.045143  184.123596      5.082251
35  trial_val_single_df  europarl       Q4      0.367-0.611     65  32.892308
    ↪      31.0        5.0       80.0  18.193511  196.830769      5.001426
```

[ ]:
```
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
  Downloading 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)
                              400.7/400.7
MB 2.5 MB/s eta 0:00:00
Installing collected packages: en-core-web-lg
Successfully installed en-core-web-lg-3.8.0
  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.
```

[ ]:
```
nlp = spacy.load("en_core_web_lg")
```

[ ]:
```
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
```

[ ]:
```
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
6500  Our results and the sequences we provide will …  [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
2944  had prepared for him a great room, where befor…  [AUX, VERB, ADP, PRON,␣
 ↪DET, ADJ, NOUN, PUNCT, …  [aux, ROOT, dative, pobj, det, amod, dobj, pun… ␣
 ↪[(Tense=Past, VerbForm=Fin), (Aspect=Perf, Ten…           1.301587
2024  (EL) The next item is the statements by the Co…  [PUNCT, PROPN, PUNCT,␣
 ↪DET, ADJ, NOUN, AUX, DET…  [punct, ROOT, punct, det, amod, nsubj, ROOT, d… ␣
 ↪[(PunctSide=Ini, PunctType=Brck), (Number=Sing…           1.421053

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

                               sentence_no_contractions                                ␣
 ↪             pos_sequence                                                 dep_sequence  ␣
 ↪                                     morph_sequence  morph_complexity
478  At the time I could not get a majority of Parl…  [ADP, DET, NOUN, PRON,␣
 ↪AUX, PART, VERB, DET, N…  [prep, det, pobj, nsubj, aux, neg, ccomp, det,… ␣
 ↪[(), (Definite=Def, PronType=Art), (Number=Sin…           1.363636
```

```
721  'Mr Poos is known for both his opposition to T…  [PUNCT, PROPN, PROPN, AUX,␣
 ↪VERB, ADP, CCONJ, P…  [punct, compound, nsubjpass, auxpass, ROOT, pr… ␣
 ↪[(PunctSide=Ini, PunctType=Quot), (Number=Sing…          1.521739
312  All of these findings raise many questions as …  [PRON, ADP, DET, NOUN,␣
 ↪VERB, ADJ, NOUN, ADP, A…  [nsubj, prep, det, pobj, ROOT, amod, dobj, pre… ␣
 ↪[(), (), (Number=Plur, PronType=Dem), (Number=…          0.892857


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


                              sentence_no_contractions                                ␣
 ↪            pos_sequence                                          dep_sequence    ␣
 ↪                                    morph_sequence  morph_complexity
521  The aim of the meeting will be to formalise th…  [DET, NOUN, ADP, DET,␣
 ↪NOUN, AUX, AUX, PART, VE…  [det, nsubj, prep, det, pobj, aux, ROOT, aux, … ␣
 ↪[(Definite=Def, PronType=Art), (Number=Sing), …          1.028571
737  SEM confirmed many of the observations made by…  [PROPN, VERB, ADJ, ADP,␣
 ↪DET, NOUN, VERB, ADP, …  [nsubj, ROOT, dobj, prep, det, pobj, acl, agen… ␣
 ↪[(Number=Sing), (Tense=Past, VerbForm=Fin), (D…          1.181818
740  It is a pleasure to welcome the Presidents and…  [PRON, AUX, DET, NOUN,␣
 ↪PART, VERB, DET, NOUN, …  [nsubj, ROOT, det, attr, aux, relcl, det, dobj… ␣
 ↪[(Case=Nom, Gender=Neut, Number=Sing, Person=3…          1.121212


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


                              sentence_no_contractions                                ␣
 ↪            pos_sequence                                          dep_sequence    ␣
 ↪                                    morph_sequence  morph_complexity
142  The burden of Egypt: "Behold, Yahweh rides on …  [DET, NOUN, ADP, PROPN,␣
 ↪PUNCT, PUNCT, VERB, PU…  [det, nsubj, prep, pobj, punct, punct, advcl, … ␣
 ↪[(Definite=Def, PronType=Art), (Number=Sing), …          1.250000
6    They also allow for easy compensation for the …  [PRON, ADV, VERB, ADP,␣
 ↪ADJ, NOUN, ADP, DET, NO…  [nsubj, advmod, ROOT, prep, amod, pobj, prep, … ␣
 ↪[(Case=Nom, Number=Plur, Person=3, PronType=Pr…          1.050000
97              This only records part of what I said.  [PRON, ADV, VERB, NOUN,␣
 ↪ADP, PRON, PRON, VERB,…  [nsubj, advmod, ROOT, dobj, prep, dobj, nsubj,… ␣
 ↪[(Number=Sing, PronType=Dem), (), (Number=Sing…          1.555556


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

```
                                sentence_no_contractions                                ␣
 ↪            pos_sequence                                          dep_sequence    ␣
 ↪                                    morph_sequence  morph_complexity
521  On the ninth day of the fourth month the famin…  [ADP, DET, ADJ, NOUN, ADP,␣
 ↪DET, ADJ, NOUN, DET…  [prep, det, amod, pobj, prep, det, amod, pobj,…  [(),␣
 ↪(Definite=Def, PronType=Art), (Degree=Pos…           1.172414
737  Unfortunately, efforts to characterize cogniti…  [ADV, PUNCT, NOUN, PART,␣
 ↪VERB, ADJ, NOUN, AUX,…  [advmod, punct, nsubjpass, aux, acl, amod, dob…  [(),␣
 ↪(PunctType=Comm), (Number=Plur), (), (Ver…           1.000000
740  For many years, the EU fleet has suffered from…  [ADP, ADJ, NOUN, PUNCT,␣
 ↪DET, PROPN, NOUN, AUX,…  [prep, amod, pobj, punct, det, compound, nsubj…  [(),␣
 ↪(Degree=Pos), (Number=Plur), (PunctType=C…           1.428571


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


                                sentence_no_contractions                                ␣
 ↪            pos_sequence                                          dep_sequence    ␣
 ↪                                    morph_sequence  morph_complexity
142  To determine if this is due to different metho…  [PART, VERB, SCONJ, PRON,␣
 ↪AUX, ADJ, ADP, ADJ, …  [aux, advcl, mark, nsubj, ccomp, acomp, prep, …  [(),␣
 ↪(VerbForm=Inf), (), (Number=Sing, PronTyp…           1.166667
6    What plans does the Commission have to introdu…  [PRON, VERB, AUX, DET,␣
 ↪PROPN, VERB, PART, VERB…  [nsubj, csubj, aux, det, nsubj, ROOT, aux, xco… ␣
 ↪[(), (Number=Sing, Person=3, Tense=Pres, VerbF…          1.411765
97   Unfortunately, ETA has once again revealed its…  [ADV, PUNCT, PROPN, AUX,␣
 ↪ADV, ADV, VERB, PRON,…  [advmod, punct, nsubj, aux, advmod, advmod, RO…  [(),␣
 ↪(PunctType=Comm), (Number=Sing), (Mood=In…           1.425000
```

[ ]:
```
# 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: 7000 entries, 0 to 6999
Data columns (total 12 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        7000 non-null   object
 1   corpus                    7000 non-null   object
 2   sentence                  7000 non-null   object
```

```
 3   token                     6995 non-null   object
 4   complexity                7000 non-null   float64
 5   is_duplicated             7000 non-null   object
 6   sentence_no_contractions  7000 non-null   object
 7   contraction_expanded      7000 non-null   bool
 8   pos_sequence              7000 non-null   object
 9   dep_sequence              7000 non-null   object
 10  morph_sequence            7000 non-null   object
 11  morph_complexity          7000 non-null   float64
dtypes: bool(1), float64(2), object(9)
memory usage: 608.5+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1300 entries, 0 to 1299
Data columns (total 12 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        1300 non-null   object
 1   corpus                    1300 non-null   object
 2   sentence                  1300 non-null   object
 3   token                     1300 non-null   object
 4   complexity                1300 non-null   float64
 5   is_duplicated             1300 non-null   object
 6   sentence_no_contractions  1300 non-null   object
 7   contraction_expanded      1300 non-null   bool
 8   pos_sequence              1300 non-null   object
 9   dep_sequence              1300 non-null   object
 10  morph_sequence            1300 non-null   object
 11  morph_complexity          1300 non-null   float64
dtypes: bool(1), float64(2), object(9)
memory usage: 113.1+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 12 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        1000 non-null   object
 1   corpus                    1000 non-null   object
 2   sentence                  1000 non-null   object
 3   token                     998 non-null    object
 4   complexity                1000 non-null   float64
 5   is_duplicated             1000 non-null   object
 6   sentence_no_contractions  1000 non-null   object
 7   contraction_expanded      1000 non-null   bool
 8   pos_sequence              1000 non-null   object
 9   dep_sequence              1000 non-null   object
 10  morph_sequence            1000 non-null   object
```

```
 11  morph_complexity          1000 non-null   float64
dtypes: bool(1), float64(2), object(9)
memory usage: 87.0+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 12 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        250 non-null    object
 1   corpus                    250 non-null    object
 2   sentence                  250 non-null    object
 3   token                     250 non-null    object
 4   complexity                250 non-null    float64
 5   is_duplicated             250 non-null    object
 6   sentence_no_contractions  250 non-null    object
 7   contraction_expanded      250 non-null    bool
 8   pos_sequence              250 non-null    object
 9   dep_sequence              250 non-null    object
 10  morph_sequence            250 non-null    object
 11  morph_complexity          250 non-null    float64
dtypes: bool(1), float64(2), object(9)
memory usage: 21.9+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 12 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        1000 non-null   object
 1   corpus                    1000 non-null   object
 2   sentence                  1000 non-null   object
 3   token                     1000 non-null   object
 4   complexity                1000 non-null   float64
 5   is_duplicated             1000 non-null   object
 6   sentence_no_contractions  1000 non-null   object
 7   contraction_expanded      1000 non-null   bool
 8   pos_sequence              1000 non-null   object
 9   dep_sequence              1000 non-null   object
 10  morph_sequence            1000 non-null   object
 11  morph_complexity          1000 non-null   float64
dtypes: bool(1), float64(2), object(9)
memory usage: 87.0+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 12 columns):
 #   Column                    Non-Null Count  Dtype
```

```
---  ------                     --------------  -----
 0   id                         250 non-null    object
 1   corpus                     250 non-null    object
 2   sentence                   250 non-null    object
 3   token                      250 non-null    object
 4   complexity                 250 non-null    float64
 5   is_duplicated              250 non-null    object
 6   sentence_no_contractions   250 non-null    object
 7   contraction_expanded       250 non-null    bool
 8   pos_sequence               250 non-null    object
 9   dep_sequence               250 non-null    object
 10  morph_sequence             250 non-null    object
 11  morph_complexity           250 non-null    float64
dtypes: bool(1), float64(2), object(9)
memory usage: 21.9+ 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): 0.4166666666666666

Sample rows from train_single_df:
                              id  complexity  binary_complexity
0  3IQ9O0AYW6ZPOAQ7VNRXLNM4D1DITZ    0.250000                  0
1  3PA41K45VN4U7YG4VFEGPOVYAII7PP    0.382353                  1
2  36818Z1KV3D5JB9F4KTTMCUN6U7A3I    0.328947                  1
3  3VJ4PFXFJ37PI5MYJ4PU9LKNJ9SUAF    0.315789                  1
4  37AQKJ12TXOFXO6IPZQ1ZUODOJPTTP    0.267857                  0

Sample rows from train_multi_df:
                              id  complexity  binary_complexity
0  3T2EL38U0MK9MPNAD5X3JSYWH8BQXO    0.600000                  1
1  388CL5C1RJN1927IGW7LZKB8JDSLHQ    0.328947                  0
2  3A3KKYU7P3H3CAKSB7U0000KY58MW4    0.350000                  0
3  3FBEFUUYRK54GUWXNMRRTF67GLFA6U    0.638889                  1
4  36QZ6V1589DTI18S04BLULET5D3SU9    0.116667                  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: 7000 entries, 0 to 6999
Data columns (total 13 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
```

```
 0   id                         7000 non-null   object
 1   corpus                     7000 non-null   object
 2   sentence                   7000 non-null   object
 3   token                      6995 non-null   object
 4   complexity                 7000 non-null   float64
 5   is_duplicated              7000 non-null   object
 6   sentence_no_contractions   7000 non-null   object
 7   contraction_expanded       7000 non-null   bool
 8   pos_sequence               7000 non-null   object
 9   dep_sequence               7000 non-null   object
 10  morph_sequence             7000 non-null   object
 11  morph_complexity           7000 non-null   float64
 12  binary_complexity          7000 non-null   int64
dtypes: bool(1), float64(2), int64(1), object(9)
memory usage: 663.2+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1300 entries, 0 to 1299
Data columns (total 13 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   id                         1300 non-null   object
 1   corpus                     1300 non-null   object
 2   sentence                   1300 non-null   object
 3   token                      1300 non-null   object
 4   complexity                 1300 non-null   float64
 5   is_duplicated              1300 non-null   object
 6   sentence_no_contractions   1300 non-null   object
 7   contraction_expanded       1300 non-null   bool
 8   pos_sequence               1300 non-null   object
 9   dep_sequence               1300 non-null   object
 10  morph_sequence             1300 non-null   object
 11  morph_complexity           1300 non-null   float64
 12  binary_complexity          1300 non-null   int64
dtypes: bool(1), float64(2), int64(1), object(9)
memory usage: 123.3+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 13 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   id                         1000 non-null   object
 1   corpus                     1000 non-null   object
 2   sentence                   1000 non-null   object
 3   token                      998 non-null    object
 4   complexity                 1000 non-null   float64
 5   is_duplicated              1000 non-null   object
```

```
 6   sentence_no_contractions  1000 non-null   object
 7   contraction_expanded      1000 non-null   bool
 8   pos_sequence              1000 non-null   object
 9   dep_sequence              1000 non-null   object
 10  morph_sequence            1000 non-null   object
 11  morph_complexity          1000 non-null   float64
 12  binary_complexity         1000 non-null   int64
dtypes: bool(1), float64(2), int64(1), object(9)
memory usage: 94.9+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 13 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        250 non-null    object
 1   corpus                    250 non-null    object
 2   sentence                  250 non-null    object
 3   token                     250 non-null    object
 4   complexity                250 non-null    float64
 5   is_duplicated             250 non-null    object
 6   sentence_no_contractions  250 non-null    object
 7   contraction_expanded      250 non-null    bool
 8   pos_sequence              250 non-null    object
 9   dep_sequence              250 non-null    object
 10  morph_sequence            250 non-null    object
 11  morph_complexity          250 non-null    float64
 12  binary_complexity         250 non-null    int64
dtypes: bool(1), float64(2), int64(1), object(9)
memory usage: 23.8+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 13 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        1000 non-null   object
 1   corpus                    1000 non-null   object
 2   sentence                  1000 non-null   object
 3   token                     1000 non-null   object
 4   complexity                1000 non-null   float64
 5   is_duplicated             1000 non-null   object
 6   sentence_no_contractions  1000 non-null   object
 7   contraction_expanded      1000 non-null   bool
 8   pos_sequence              1000 non-null   object
 9   dep_sequence              1000 non-null   object
 10  morph_sequence            1000 non-null   object
 11  morph_complexity          1000 non-null   float64
```

```
 12  binary_complexity         1000 non-null   int64
dtypes: bool(1), float64(2), int64(1), object(9)
memory usage: 94.9+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 13 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        250 non-null    object
 1   corpus                    250 non-null    object
 2   sentence                  250 non-null    object
 3   token                     250 non-null    object
 4   complexity                250 non-null    float64
 5   is_duplicated             250 non-null    object
 6   sentence_no_contractions  250 non-null    object
 7   contraction_expanded      250 non-null    bool
 8   pos_sequence              250 non-null    object
 9   dep_sequence              250 non-null    object
 10  morph_sequence            250 non-null    object
 11  morph_complexity          250 non-null    float64
 12  binary_complexity         250 non-null    int64
dtypes: bool(1), float64(2), int64(1), object(9)
memory usage: 23.8+ 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 is_duplicated
sentence_no_contractions  contraction_expanded
pos_sequence                                        dep_sequence
morph_sequence  morph_complexity  binary_complexity
0  3IQ9O0AYW6ZPOAQ7VNRXLNM4D1DITZ    biomed  The development of sexually
dimorphic reproduc…           organs    0.250000            {}  The
development of sexually dimorphic reproduc…                 False  [DET, NOUN,
ADP, ADV, ADJ, ADJ, NOUN, AUX, DET…  [det, nsubj, prep, advmod, amod, amod,
pobj, R…  [(Definite=Def, PronType=Art), (Number=Sing), …          1.200000
0
1  3PA41K45VN4U7YG4VFEGPOVYAII7PP    biomed  We find that the majority of the
olfactory rec…            usage    0.382353            {}  We find that the
majority of the olfactory rec…                 False  [PRON, VERB, SCONJ, DET,
NOUN, ADP, DET, ADJ, …  [nsubj, ROOT, mark, det, nsubjpass, prep, det,…
```

```
[(Case=Nom, Number=Plur, Person=1, PronType=Pr…          1.142857
1
2  36818Z1KV3D5JB9F4KTTMCUN6U7A3I     bible  His lord was angry, and delivered
him to the t…       tormentors    0.328947           {}  His lord was angry,
and delivered him to the t…                 False  [PRON, NOUN, AUX, ADJ,
PUNCT, CCONJ, VERB, PRO…  [poss, nsubj, ROOT, acomp, punct, cc, conj, do…
[(Gender=Masc, Number=Sing, Person=3, Poss=Yes…          1.956522
1
3  3VJ4PFXFJ37PI5MYJ4PU9LKNJ9SUAF  europarl  The Taiwanese Government has
informed the Coun…  representations    0.315789           {}  The Taiwanese
Government has informed the Coun…                 False  [DET, ADJ, PROPN,
AUX, VERB, DET, PROPN, PUNCT…  [det, amod, nsubj, aux, ROOT, det, dobj,
punct…  [(Definite=Def, PronType=Art), (Degree=Pos), (…          1.432432
1
4  37AQKJ12TXOFXO6IPZQ1ZUODOJPTTP  europarl  However, I too want to thank
everyone who took…         relation    0.267857           {}  However, I too
want to thank everyone who took…                 False  [ADV, PUNCT, PRON,
ADV, VERB, PART, VERB, PRON…  [advmod, punct, nsubj, advmod, ROOT, aux,
xcom…  [(), (PunctType=Comm), (Case=Nom, Number=Sing,…          1.156250
0
                              id    corpus
sentence                      token  complexity is_duplicated
sentence_no_contractions  contraction_expanded
pos_sequence                                        dep_sequence
morph_sequence  morph_complexity  binary_complexity
0  3T2EL38U0MK9MPNAD5X3JSYWH8BQXO    biomed  CA = chronic arthritis; CIA =
collagen-induced…      rheumatoid arthritis    0.600000           {}  CA =
chronic arthritis; CIA = collagen-induced…                 False  [PROPN, ADP,
ADJ, NOUN, PUNCT, PROPN, PUNCT, N…  [nmod, punct, amod, ROOT, punct, nmod,
punct, …  [(Number=Sing), (), (Degree=Pos), (Number=Sing…          0.857143
1
1  388CL5C1RJN1927IGW7LZKB8JDSLHQ  europarl  Appointments to parliamentary
committees (vote…  parliamentary committees    0.328947           {}
Appointments to parliamentary committees (vote…                 False  [NOUN,
ADP, ADJ, NOUN, PUNCT, VERB, PUNCT, VER…  [nsubj, prep, amod, pobj, punct,
ccomp, punct,…  [(Number=Plur), (), (Degree=Pos), (Number=Plur…
0.888889                   0
2  3A3KKYU7P3H3CAKSB7U0000KY58MW4    biomed  The HG9 strain represents a major
epistasis-ba…               mouse model    0.350000           {}  The HG9
strain represents a major epistasis-ba…                 False  [DET, PROPN,
NOUN, VERB, DET, ADJ, NOUN, PUNCT…  [det, compound, nsubj, ROOT, det, amod,
npadvm…  [(Definite=Def, PronType=Art), (Number=Sing), …          1.093750
0
3  3FBEFUUYRK54GUWXNMRRTF67GLFA6U     bible  For there is an annulling of a
foregoing comma…     foregoing commandment    0.638889           {}  For
there is an annulling of a foregoing comma…                 False  [ADP, PRON,
VERB, DET, NOUN, ADP, DET, NOUN, N…  [prep, expl, ROOT, det, attr, prep, det,
compo…  [(), (), (Mood=Ind, Number=Sing, Person=3, Ten…          1.333333
```

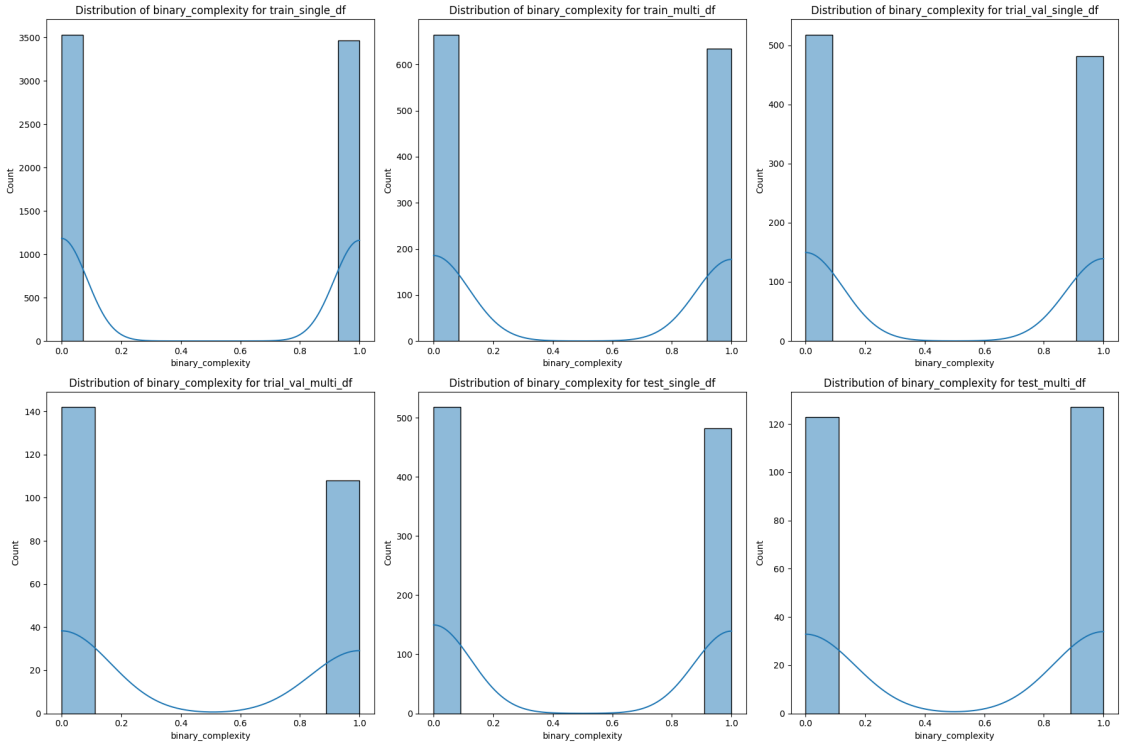
```
1
4  36QZ6V1589DTI18S04BLULET5D3SU9     bible  Ezra the priest, with certain heads
of fathers…                  first day    0.116667            {}  Ezra the
priest, with certain heads of fathers…                  False  [PROPN, DET,
NOUN, PUNCT, ADP, ADJ, NOUN, ADP,…  [nsubjpass, det, appos, punct, prep, amod,
pob…  [(Number=Sing), (Definite=Def, PronType=Art), …          1.148936
0
                              id  corpus
sentence    token  complexity is_duplicated
sentence_no_contractions  contraction_expanded
pos_sequence                                          dep_sequence
morph_sequence  morph_complexity  binary_complexity
0  3ZQA3IO31BRYBCP1RZKSZEZVXRG1OZ  biomed  In addition to colorectal neoplams,
these indi…  pigment    0.350000            {}  In addition to colorectal
neoplams, these indi…                  False  [ADP, NOUN, ADP, ADJ, NOUN,
PUNCT, DET, NOUN, …  [prep, pobj, prep, amod, pobj, punct, det, nsu…  [(),
(Number=Sing), (), (Degree=Pos), (Number=…          1.050847
1
1  3Z3R5YC0P3N5EJOHUFLCIQ9CX7PTFJ   bible  The Queen of the South will rise up
in the jud…     ends    0.302632            {}  The Queen of the South will
rise up in the jud…                  False  [DET, PROPN, ADP, DET, PROPN, AUX,
VERB, ADP, …  [det, nsubj, prep, det, pobj, aux, ROOT, prt, …
[(Definite=Def, PronType=Art), (Number=Sing), …          1.142857
1
2  3URJ6VVYUPNF3BMKEH3UXC6Y9BQ4OF  biomed  Since the parental strains differ in
susceptib…    class    0.261905            {}  Since the parental strains
differ in susceptib…                   False  [SCONJ, DET, ADJ, NOUN, VERB, ADP,
NOUN, ADP, …  [mark, det, amod, nsubj, advcl, prep, pobj, pr…  [(),
(Definite=Def, PronType=Art), (Degree=Pos…          1.073171
0
3  3MVY4USGB6NO9ADS6NM7BIQIBGKSI1   bible  For the judgment is against you; for
you have …    Tabor    0.633333            {}  For the judgment is against
you; for you have …                  False  [ADP, DET, NOUN, AUX, ADP, PRON,
PUNCT, SCONJ,…  [prep, det, pobj, ccomp, prep, pobj, punct, ma…  [(),
(Definite=Def, PronType=Art), (Number=Sin…          1.347826
1
4  30U1YOGZGAW71ZX6E9LWKLA5JD8SDZ   bible  having a great and high wall; having
twelve ga…   tribes    0.175000            {}  having a great and high wall;
having twelve ga…                 False  [VERB, DET, ADJ, CCONJ, ADJ, NOUN,
PUNCT, VERB…  [ROOT, det, amod, cc, conj, dobj, punct, conj,…
[(Aspect=Prog, Tense=Pres, VerbForm=Part), (De…          1.236842
0
                              id    corpus
sentence                token  complexity is_duplicated
sentence_no_contractions  contraction_expanded
pos_sequence                                          dep_sequence
morph_sequence  morph_complexity  binary_complexity
0  3D17ECOUOEV9PNWF8100BB1K20731T     bible  But some of the itinerant Jews,
```

```
exorcists, too…       itinerant Jews    0.600000            {}  But some of
the itinerant Jews, exorcists, too…                  False  [CCONJ, PRON, ADP,
DET, ADJ, PROPN, PUNCT, NOU…  [cc, nsubj, prep, det, amod, pobj, punct,
appo…  [(ConjType=Cmp), (), (), (Definite=Def, PronTy…          1.365854
1
1  3XBXDSS888JYVS7XL0P726Z273BLXJ  europarl  The next item is the report by
Esther de Lange…       EU legislation    0.285714            {}  The next item
is the report by Esther de Lange…                  False  [DET, ADJ, NOUN, AUX,
DET, NOUN, ADP, PROPN, P…  [det, amod, nsubj, ROOT, det, attr, prep, comp…
[(Definite=Def, PronType=Art), (Degree=Pos), (…          1.102564
0
2  3GITHABACYLNIC7L9OKTP89VZONN2N    biomed  Alternatively, the unusual
transcriptional reg…  olfactory receptors    0.725000            {}
Alternatively, the unusual transcriptional reg…                  False  [ADV,
PUNCT, DET, ADJ, ADJ, NOUN, ADP, ADJ, NO…  [advmod, punct, det, amod, amod,
nsubj, prep, …  [(), (PunctType=Comm), (Definite=Def, PronType…
1.260870                  1
3  31MCUE39BKM6T2MIQKL3IY5Q4Q13G6    biomed  Genetic disruption of the Dhcr7
results in neo…   neonatal lethality    0.547619            {}  Genetic
disruption of the Dhcr7 results in neo…                  False  [ADJ, NOUN,
ADP, DET, PROPN, NOUN, ADP, ADJ, N…  [amod, ROOT, prep, det, compound, pobj,
prep, …  [(Degree=Pos), (Number=Sing), (), (Definite=De…          0.923077
1
4  37PGLWGSJT6QLR0K1ED5KWZ8UO3IKA     bible  In it you shall not sow, neither
reap that whi…      undressed vines    0.525000            {}  In it you shall
not sow, neither reap that whi…                  False  [ADP, PRON, PRON, AUX,
PART, VERB, PUNCT, CCON…  [prep, pobj, nsubj, aux, neg, ROOT, punct, pre…
[(), (Case=Acc, Gender=Neut, Number=Sing, Pers…          1.500000
1
                               id    corpus
sentence         token  complexity is_duplicated
sentence_no_contractions  contraction_expanded
pos_sequence                                          dep_sequence
morph_sequence  morph_complexity  binary_complexity
0  3ZURAPD288N45ZC8SW12CKQH5QPF1R    biomed  We show that in p150CAF-1-depleted
ES cells, w…  perturbation    0.484375            {}  We show that in
p150CAF-1-depleted ES cells, w…                  False  [PRON, VERB, SCONJ,
ADP, ADV, PUNCT, VERB, NOU…  [nsubj, ROOT, mark, prep, npadvmod, punct, amo…
[(Case=Nom, Number=Plur, Person=1, PronType=Pr…          1.133333
1
1  36D1BWBEHN1H0UMLXN5TCTKVUXL2M8    biomed  Lung development is a complex
process that inv…       process    0.250000            {}  Lung development is
a complex process that inv…                  False  [PROPN, NOUN, AUX, DET,
ADJ, NOUN, PRON, VERB,…  [compound, nsubj, ROOT, det, amod, attr, nsubj…
[(Number=Sing), (Number=Sing), (Mood=Ind, Numb…          1.407407
0
2  3QX22DUVOOHQXLKNLXP4EYH6RZBVME  europarl  That is why we want to introduce
the role of m…          role    0.050000            {}  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
3  3HXCEECSQMT7OMEB5X2ITZH9OICZYL  europarl  (CS) I would just like to emphasise
that this …     groupings    0.210526           {}  (CS) I would just like to
emphasise that this …                 False  [PUNCT, PROPN, PUNCT, PRON, AUX,
ADV, VERB, PA…  [punct, npadvmod, punct, nsubj, aux, advmod, R…
[(PunctSide=Ini, PunctType=Brck), (Number=Sing…          1.254545
0
4  3WGCNLZJKF877FYC1Q6COKNWTFRD10  europarl  I am from a border county myself
and I am a re…       process    0.183333           {}  I am from a border
county myself and I am a re…                 False  [PRON, AUX, ADP, DET,
NOUN, NOUN, PRON, CCONJ,…  [nsubj, ROOT, prep, det, compound, pobj, npadv…
[(Case=Nom, Number=Sing, Person=1, PronType=Pr…          1.609756
0
                              id    corpus
sentence                  token  complexity is_duplicated
sentence_no_contractions  contraction_expanded
pos_sequence                                        dep_sequence
morph_sequence  morph_complexity  binary_complexity
0  3FK4G712NX0D3OGOBZGLFKW5KGISST     bible  He shall put no oil on it, neither
shall he pu…          sin offering    0.450000           {}  He shall put no
oil on it, neither shall he pu…                 False  [PRON, AUX, VERB, DET,
NOUN, ADP, PRON, PUNCT,…  [nsubj, aux, ROOT, det, dobj, prep, pobj, punc…
[(Case=Nom, Gender=Masc, Number=Sing, Person=3…          1.833333
1
1  3UQVX1UPFSHKXGFE8IIVEWDIRVC02P    biomed  During the last few years the
Wnt1-Cre transge…         powerful tool    0.305556           {}  During the
last few years the Wnt1-Cre transge…                  False  [ADP, DET, ADJ,
ADJ, NOUN, DET, NUM, PUNCT, NO…  [prep, det, amod, amod, pobj, det, compound,
p…  [(), (Definite=Def, PronType=Art), (Degree=Pos…          1.161290
0
2  3T2EL38U0MK9MPNAD5X3JSYWH9XXQJ  europarl  The next item is the report by Mrs
Fajon, on b…      external borders    0.343750           {}  The next item is
the report by Mrs Fajon, on b…                  False  [DET, ADJ, NOUN, AUX,
DET, NOUN, ADP, PROPN, P…  [det, amod, nsubj, ROOT, det, attr, prep, comp…
[(Definite=Def, PronType=Art), (Degree=Pos), (…          1.137500
0
3  37AQKJ12TXOFXO6IPZQ1ZUODOJMTTM    biomed  The pathogenesis and developmental
relationshi…  pulmonary hypoplasia    0.675000           {}  The pathogenesis
and developmental relationshi…                  False  [DET, NOUN, CCONJ, ADJ,
NOUN, ADP, ADJ, NOUN, …  [det, nsubjpass, cc, amod, conj, prep, amod, p…
[(Definite=Def, PronType=Art), (Number=Sing), …          1.400000
1
4  3NZ1E5QA6Z1DG01BOHHIWKCD28P5B4     bible  Moreover I will make a covenant of
peace with …  everlasting covenant    0.444444           {}  Moreover I will
make a covenant of peace with …                 False  [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
1
```

[ ]:
```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()
```

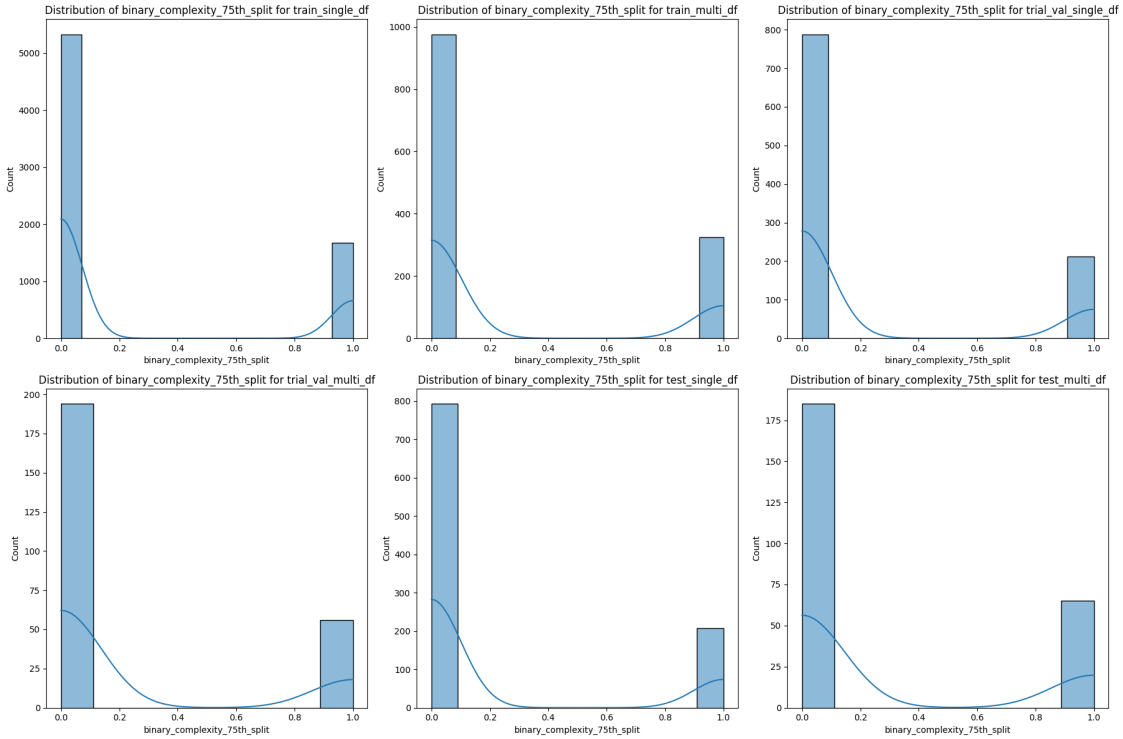


[ ]:
```
train_single_75th = train_single_df['complexity'].quantile(0.75)
train_multi_75th = train_multi_df['complexity'].quantile(0.75)

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    5326
1    1674
Name: count, dtype: int64

Distribution of 'binary_complexity_75th_split' in train_multi_df:
binary_complexity_75th_split
0    976
1    324
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: 7000 entries, 0 to 6999
Data columns (total 14 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   id                             7000 non-null   object
 1   corpus                         7000 non-null   object
 2   sentence                       7000 non-null   object
 3   token                          6995 non-null   object
 4   complexity                     7000 non-null   float64
 5   is_duplicated                  7000 non-null   object
 6   sentence_no_contractions       7000 non-null   object
 7   contraction_expanded           7000 non-null   bool
 8   pos_sequence                   7000 non-null   object
 9   dep_sequence                   7000 non-null   object
 10  morph_sequence                 7000 non-null   object
 11  morph_complexity               7000 non-null   float64
 12  binary_complexity              7000 non-null   int64
 13  binary_complexity_75th_split   7000 non-null   int64
dtypes: bool(1), float64(2), int64(2), object(9)
memory usage: 717.9+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1300 entries, 0 to 1299
Data columns (total 14 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   id                             1300 non-null   object
 1   corpus                         1300 non-null   object
 2   sentence                       1300 non-null   object
 3   token                          1300 non-null   object
 4   complexity                     1300 non-null   float64
 5   is_duplicated                  1300 non-null   object
 6   sentence_no_contractions       1300 non-null   object
 7   contraction_expanded           1300 non-null   bool
 8   pos_sequence                   1300 non-null   object
 9   dep_sequence                   1300 non-null   object
 10  morph_sequence                 1300 non-null   object
 11  morph_complexity               1300 non-null   float64
 12  binary_complexity              1300 non-null   int64
 13  binary_complexity_75th_split   1300 non-null   int64
```

```
dtypes: bool(1), float64(2), int64(2), object(9)
memory usage: 133.4+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 14 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   id                             1000 non-null   object
 1   corpus                         1000 non-null   object
 2   sentence                       1000 non-null   object
 3   token                          998 non-null    object
 4   complexity                     1000 non-null   float64
 5   is_duplicated                  1000 non-null   object
 6   sentence_no_contractions       1000 non-null   object
 7   contraction_expanded           1000 non-null   bool
 8   pos_sequence                   1000 non-null   object
 9   dep_sequence                   1000 non-null   object
 10  morph_sequence                 1000 non-null   object
 11  morph_complexity               1000 non-null   float64
 12  binary_complexity              1000 non-null   int64
 13  binary_complexity_75th_split   1000 non-null   int64
dtypes: bool(1), float64(2), int64(2), object(9)
memory usage: 102.7+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 14 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   id                             250 non-null    object
 1   corpus                         250 non-null    object
 2   sentence                       250 non-null    object
 3   token                          250 non-null    object
 4   complexity                     250 non-null    float64
 5   is_duplicated                  250 non-null    object
 6   sentence_no_contractions       250 non-null    object
 7   contraction_expanded           250 non-null    bool
 8   pos_sequence                   250 non-null    object
 9   dep_sequence                   250 non-null    object
 10  morph_sequence                 250 non-null    object
 11  morph_complexity               250 non-null    float64
 12  binary_complexity              250 non-null    int64
 13  binary_complexity_75th_split   250 non-null    int64
dtypes: bool(1), float64(2), int64(2), object(9)
memory usage: 25.8+ KB
None
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 14 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              1000 non-null   object
 1   corpus                          1000 non-null   object
 2   sentence                        1000 non-null   object
 3   token                           1000 non-null   object
 4   complexity                      1000 non-null   float64
 5   is_duplicated                   1000 non-null   object
 6   sentence_no_contractions        1000 non-null   object
 7   contraction_expanded            1000 non-null   bool
 8   pos_sequence                    1000 non-null   object
 9   dep_sequence                    1000 non-null   object
 10  morph_sequence                  1000 non-null   object
 11  morph_complexity                1000 non-null   float64
 12  binary_complexity               1000 non-null   int64
 13  binary_complexity_75th_split    1000 non-null   int64
dtypes: bool(1), float64(2), int64(2), object(9)
memory usage: 102.7+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 14 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              250 non-null    object
 1   corpus                          250 non-null    object
 2   sentence                        250 non-null    object
 3   token                           250 non-null    object
 4   complexity                      250 non-null    float64
 5   is_duplicated                   250 non-null    object
 6   sentence_no_contractions        250 non-null    object
 7   contraction_expanded            250 non-null    bool
 8   pos_sequence                    250 non-null    object
 9   dep_sequence                    250 non-null    object
 10  morph_sequence                  250 non-null    object
 11  morph_complexity                250 non-null    float64
 12  binary_complexity               250 non-null    int64
 13  binary_complexity_75th_split    250 non-null    int64
dtypes: bool(1), float64(2), int64(2), object(9)
memory usage: 25.8+ 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 is_duplicated
sentence_no_contractions  contraction_expanded
pos_sequence                                          dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split
0  3IQ9OOAYW6ZPOAQ7VNRXLNM4D1DITZ    biomed  The development of sexually
dimorphic reproduc…           organs    0.250000            {}  The
development of sexually dimorphic reproduc…                  False  [DET, NOUN,
ADP, ADV, ADJ, ADJ, NOUN, AUX, DET…  [det, nsubj, prep, advmod, amod, amod,
pobj, R…  [(Definite=Def, PronType=Art), (Number=Sing), …          1.200000
0                             0
1  3PA41K45VN4U7YG4VFEGPOVYAII7PP    biomed  We find that the majority of the
olfactory rec…            usage    0.382353            {}  We find that the
majority of the olfactory rec…                 False  [PRON, VERB, SCONJ, DET,
NOUN, ADP, DET, ADJ, …  [nsubj, ROOT, mark, det, nsubjpass, prep, det,…
[(Case=Nom, Number=Plur, Person=1, PronType=Pr…          1.142857
1                             1
2  36818Z1KV3D5JB9F4KTTMCUN6U7A3I     bible  His lord was angry, and delivered
him to the t…       tormentors    0.328947            {}  His lord was angry,
and delivered him to the t…                  False  [PRON, NOUN, AUX, ADJ,
PUNCT, CCONJ, VERB, PRO…  [poss, nsubj, ROOT, acomp, punct, cc, conj, do…
[(Gender=Masc, Number=Sing, Person=3, Poss=Yes…          1.956522
1                             0
3  3VJ4PFXFJ37PI5MYJ4PU9LKNJ9SUAF  europarl  The Taiwanese Government has
informed the Coun…  representations    0.315789            {}  The Taiwanese
Government has informed the Coun…                   False  [DET, ADJ, PROPN,
AUX, VERB, DET, PROPN, PUNCT…  [det, amod, nsubj, aux, ROOT, det, dobj,
punct…  [(Definite=Def, PronType=Art), (Degree=Pos), (…          1.432432
1                             0
4  37AQKJ12TXOFXO6IPZQ1ZUODOJPTTP  europarl  However, I too want to thank
everyone who took…         relation    0.267857            {}  However, I too
want to thank everyone who took…                 False  [ADV, PUNCT, PRON,
ADV, VERB, PART, VERB, PRON…  [advmod, punct, nsubj, advmod, ROOT, aux,
xcom…  [(), (PunctType=Comm), (Case=Nom, Number=Sing,…          1.156250
0                             0
                              id    corpus
sentence                     token  complexity is_duplicated
sentence_no_contractions  contraction_expanded
pos_sequence                                          dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split
0  3T2EL38UOMK9MPNAD5X3JSYWH8BQXO    biomed  CA = chronic arthritis; CIA =
```

```
collagen-induced…      rheumatoid arthritis    0.600000            {}  CA =
chronic arthritis; CIA = collagen-induced…                  False  [PROPN, ADP,
ADJ, NOUN, PUNCT, PROPN, PUNCT, N…  [nmod, punct, amod, ROOT, punct, nmod,
punct, …  [(Number=Sing), (), (Degree=Pos), (Number=Sing…          0.857143
1                             1
1  388CL5C1RJN1927IGW7LZKB8JDSLHQ  europarl  Appointments to parliamentary
committees (vote…  parliamentary committees    0.328947            {}
Appointments to parliamentary committees (vote…                 False  [NOUN,
ADP, ADJ, NOUN, PUNCT, VERB, PUNCT, VER…  [nsubj, prep, amod, pobj, punct,
ccomp, punct,…  [(Number=Plur), (), (Degree=Pos), (Number=Plur…
0.888889                  0                             0
2  3A3KKYU7P3H3CAKSB7U0000KY58MW4    biomed  The HG9 strain represents a major
epistasis-ba…               mouse model    0.350000            {}  The HG9
strain represents a major epistasis-ba…                 False  [DET, PROPN,
NOUN, VERB, DET, ADJ, NOUN, PUNCT…  [det, compound, nsubj, ROOT, det, amod,
npadvm…  [(Definite=Def, PronType=Art), (Number=Sing), …          1.093750
0                             0
3  3FBEFUUYRK54GUWXNMRRTF67GLFA6U     bible  For there is an annulling of a
foregoing comma…     foregoing commandment    0.638889            {}  For
there is an annulling of a foregoing comma…                 False  [ADP, PRON,
VERB, DET, NOUN, ADP, DET, NOUN, N…  [prep, expl, ROOT, det, attr, prep, det,
compo…  [(), (), (Mood=Ind, Number=Sing, Person=3, Ten…          1.333333
1                             1
4  36QZ6V1589DTI18SO4BLULET5D3SU9     bible  Ezra the priest, with certain heads
of fathers…                 first day    0.116667            {}  Ezra the
priest, with certain heads of fathers…                  False  [PROPN, DET,
NOUN, PUNCT, ADP, ADJ, NOUN, ADP,…  [nsubjpass, det, appos, punct, prep, amod,
pob…  [(Number=Sing), (Definite=Def, PronType=Art), …          1.148936
0                             0
                              id  corpus
sentence    token  complexity is_duplicated
sentence_no_contractions  contraction_expanded
pos_sequence                                          dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split
0  3ZQA3IO31BRYBCP1RZKSZEZVXRG1OZ  biomed  In addition to colorectal neoplams,
these indi…  pigment    0.350000            {}  In addition to colorectal
neoplams, these indi…                 False  [ADP, NOUN, ADP, ADJ, NOUN,
PUNCT, DET, NOUN, …  [prep, pobj, prep, amod, pobj, punct, det, nsu…  [(),
(Number=Sing), (), (Degree=Pos), (Number=…          1.050847
1                             0
1  3Z3R5YC0P3N5EJOHUFLCIQ9CX7PTFJ   bible  The Queen of the South will rise up
in the jud…     ends    0.302632            {}  The Queen of the South will
rise up in the jud…                 False  [DET, PROPN, ADP, DET, PROPN, AUX,
VERB, ADP, …  [det, nsubj, prep, det, pobj, aux, ROOT, prt, …
[(Definite=Def, PronType=Art), (Number=Sing), …          1.142857
1                             0
2  3URJ6VVYUPNF3BMKEH3UXC6Y9BQ4OF  biomed  Since the parental strains differ in
```

```
susceptib…    class    0.261905            {}  Since the parental strains
differ in susceptib…                  False  [SCONJ, DET, ADJ, NOUN, VERB, ADP,
NOUN, ADP, …  [mark, det, amod, nsubj, advcl, prep, pobj, pr…  [(),
(Definite=Def, PronType=Art), (Degree=Pos…          1.073171
0                             0
3  3MVY4USGB6NO9ADS6NM7BIQIBGKSI1   bible  For the judgment is against you; for
you have …    Tabor    0.633333            {}  For the judgment is against
you; for you have …                   False  [ADP, DET, NOUN, AUX, ADP, PRON,
PUNCT, SCONJ,…  [prep, det, pobj, ccomp, prep, pobj, punct, ma…  [(),
(Definite=Def, PronType=Art), (Number=Sin…          1.347826
1                             1
4  30U1YOGZGAW71ZX6E9LWKLA5JD8SDZ   bible  having a great and high wall; having
twelve ga…   tribes    0.175000            {}  having a great and high wall;
having twelve ga…                 False  [VERB, DET, ADJ, CCONJ, ADJ, NOUN,
PUNCT, VERB…  [ROOT, det, amod, cc, conj, dobj, punct, conj,…
[(Aspect=Prog, Tense=Pres, VerbForm=Part), (De…          1.236842
0                             0
                              id    corpus
sentence                token  complexity is_duplicated
sentence_no_contractions  contraction_expanded
pos_sequence                                        dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split
0  3D17ECOUOEV9PNWF8100BB1K20731T     bible  But some of the itinerant Jews,
exorcists, too…       itinerant Jews    0.600000            {}  But some of
the itinerant Jews, exorcists, too…                 False  [CCONJ, PRON, ADP,
DET, ADJ, PROPN, PUNCT, NOU…  [cc, nsubj, prep, det, amod, pobj, punct,
appo…  [(ConjType=Cmp), (), (), (Definite=Def, PronTy…          1.365854
1                             1
1  3XBXDSS888JYVS7XL0P726Z273BLXJ  europarl  The next item is the report by
Esther de Lange…       EU legislation    0.285714             {}  The next item
is the report by Esther de Lange…                 False  [DET, ADJ, NOUN, AUX,
DET, NOUN, ADP, PROPN, P…  [det, amod, nsubj, ROOT, det, attr, prep, comp…
[(Definite=Def, PronType=Art), (Degree=Pos), (…          1.102564
0                             0
2  3GITHABACYLNIC7L9OKTP89VZONN2N    biomed  Alternatively, the unusual
transcriptional reg…  olfactory receptors    0.725000             {}
Alternatively, the unusual transcriptional reg…                 False  [ADV,
PUNCT, DET, ADJ, ADJ, NOUN, ADP, ADJ, NO…  [advmod, punct, det, amod, amod,
nsubj, prep, …  [(), (PunctType=Comm), (Definite=Def, PronType…
1.260870                  1                             1
3  31MCUE39BKM6T2MIQKL3IY5Q4Q13G6    biomed  Genetic disruption of the Dhcr7
results in neo…   neonatal lethality    0.547619             {}  Genetic
disruption of the Dhcr7 results in neo…                 False  [ADJ, NOUN,
ADP, DET, PROPN, NOUN, ADP, ADJ, N…  [amod, ROOT, prep, det, compound, pobj,
prep, …  [(Degree=Pos), (Number=Sing), (), (Definite=De…          0.923077
1                             1
4  37PGLWGSJT6QLR0K1ED5KWZ8UO3IKA     bible  In it you shall not sow, neither
```

```
reap that whi…      undressed vines    0.525000            {}  In it you shall
not sow, neither reap that whi…                  False  [ADP, PRON, PRON, AUX,
PART, VERB, PUNCT, CCON…  [prep, pobj, nsubj, aux, neg, ROOT, punct, pre…
[(), (Case=Acc, Gender=Neut, Number=Sing, Pers…          1.500000
1                             0
                               id    corpus
sentence         token  complexity is_duplicated
sentence_no_contractions  contraction_expanded
pos_sequence                                          dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split
0  3ZURAPD288N45ZC8SW12CKQH5QPF1R    biomed  We show that in p150CAF-1-depleted
ES cells, w…  perturbation    0.484375            {}  We show that in
p150CAF-1-depleted ES cells, w…                  False  [PRON, VERB, SCONJ,
ADP, ADV, PUNCT, VERB, NOU…  [nsubj, ROOT, mark, prep, npadvmod, punct, amo…
[(Case=Nom, Number=Plur, Person=1, PronType=Pr…          1.133333
1                             1
1  36D1BWBEHN1H0UMLXN5TCTKVUXL2M8    biomed  Lung development is a complex
process that inv…       process    0.250000            {}  Lung development is
a complex process that inv…                 False  [PROPN, NOUN, AUX, DET,
ADJ, NOUN, PRON, VERB,…  [compound, nsubj, ROOT, det, amod, attr, nsubj…
[(Number=Sing), (Number=Sing), (Mood=Ind, Numb…          1.407407
0                             0
2  3QX22DUVOOHQXLKNLXP4EYH6RZBVME  europarl  That is why we want to introduce
the role of m…          role    0.050000            {}  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
3  3HXCEECSQMT70MEB5X2ITZH9OICZYL  europarl  (CS) I would just like to emphasise
that this …     groupings    0.210526            {}  (CS) I would just like to
emphasise that this …                 False  [PUNCT, PROPN, PUNCT, PRON, AUX,
ADV, VERB, PA…  [punct, npadvmod, punct, nsubj, aux, advmod, R…
[(PunctSide=Ini, PunctType=Brck), (Number=Sing…          1.254545
0                             0
4  3WGCNLZJKF877FYC1Q6COKNWTFRD10  europarl  I am from a border county myself
and I am a re…       process    0.183333            {}  I am from a border
county myself and I am a re…                 False  [PRON, AUX, ADP, DET,
NOUN, NOUN, PRON, CCONJ,…  [nsubj, ROOT, prep, det, compound, pobj, npadv…
[(Case=Nom, Number=Sing, Person=1, PronType=Pr…          1.609756
0                             0
                               id    corpus
sentence                 token  complexity is_duplicated
sentence_no_contractions  contraction_expanded
pos_sequence                                          dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split
0  3FK4G712NX0D3OGOBZGLFKW5KGISST     bible  He shall put no oil on it, neither
```

```
shall he pu…          sin offering    0.450000            {}  He shall put no
oil on it, neither shall he pu…                  False  [PRON, AUX, VERB, DET,
NOUN, ADP, PRON, PUNCT,…  [nsubj, aux, ROOT, det, dobj, prep, pobj, punc…
[(Case=Nom, Gender=Masc, Number=Sing, Person=3…          1.833333
1                             0
1  3UQVX1UPFSHKXGFE8IIVEWDIRVC02P    biomed  During the last few years the
Wnt1-Cre transge…         powerful tool    0.305556            {}  During the
last few years the Wnt1-Cre transge…                 False  [ADP, DET, ADJ,
ADJ, NOUN, DET, NUM, PUNCT, NO…  [prep, det, amod, amod, pobj, det, compound,
p…  [(), (Definite=Def, PronType=Art), (Degree=Pos…          1.161290
0                             0
2  3T2EL38U0MK9MPNAD5X3JSYWH9XXQJ  europarl  The next item is the report by Mrs
Fajon, on b…      external borders    0.343750            {}  The next item is
the report by Mrs Fajon, on b…                  False  [DET, ADJ, NOUN, AUX,
DET, NOUN, ADP, PROPN, P…  [det, amod, nsubj, ROOT, det, attr, prep, comp…
[(Definite=Def, PronType=Art), (Degree=Pos), (…          1.137500
0                             0
3  37AQKJ12TXOFXO6IPZQ1ZUODOJMTTM    biomed  The pathogenesis and developmental
relationshi…  pulmonary hypoplasia    0.675000            {}  The pathogenesis
and developmental relationshi…                  False  [DET, NOUN, CCONJ, ADJ,
NOUN, ADP, ADJ, NOUN, …  [det, nsubjpass, cc, amod, conj, prep, amod, p…
[(Definite=Def, PronType=Art), (Number=Sing), …          1.400000
1                             1
4  3NZ1E5QA6Z1DG01BOHHIWKCD28P5B4     bible  Moreover I will make a covenant of
peace with …  everlasting covenant    0.444444            {}  Moreover I will
make a covenant of peace with …                  False  [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
1                             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    3534
1    3466
```

```
Name: count, dtype: int64

=== train_multi_df ===
binary_complexity
0    665
1    635
Name: count, dtype: int64

=== trial_val_single_df ===
binary_complexity
0    518
1    482
Name: count, dtype: int64

=== trial_val_multi_df ===
binary_complexity
0    142
1    108
Name: count, dtype: int64

=== test_single_df ===
binary_complexity
0    518
1    482
Name: count, dtype: int64

=== test_multi_df ===
binary_complexity
1    127
0    123
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
```

[ ]:
```
# 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: 7000 entries, 0 to 6999
Data columns (total 21 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              7000 non-null   object
 1   corpus                          7000 non-null   object
 2   sentence                        7000 non-null   object
 3   token                           6995 non-null   object
 4   complexity                      7000 non-null   float64
 5   is_duplicated                   7000 non-null   object
 6   sentence_no_contractions        7000 non-null   object
 7   contraction_expanded            7000 non-null   bool
 8   pos_sequence                    7000 non-null   object
 9   dep_sequence                    7000 non-null   object
 10  morph_sequence                  7000 non-null   object
 11  morph_complexity                7000 non-null   float64
 12  binary_complexity               7000 non-null   int64
 13  binary_complexity_75th_split    7000 non-null   int64
 14  snc_pos_seq                     7000 non-null   object
 15  snc_pos_alt                     7000 non-null   object
 16  snc_morph_seq                   7000 non-null   object
 17  snc_morph_alt                   7000 non-null   object
 18  snc_dep_seq                     7000 non-null   object
 19  snc_dep_alt                     7000 non-null   object
 20  snc_morph_complexity_value      7000 non-null   object
dtypes: bool(1), float64(2), int64(2), object(16)
memory usage: 1.1+ MB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1300 entries, 0 to 1299
Data columns (total 21 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              1300 non-null   object
 1   corpus                          1300 non-null   object
 2   sentence                        1300 non-null   object
 3   token                           1300 non-null   object
 4   complexity                      1300 non-null   float64
 5   is_duplicated                   1300 non-null   object
 6   sentence_no_contractions        1300 non-null   object
 7   contraction_expanded            1300 non-null   bool
 8   pos_sequence                    1300 non-null   object
 9   dep_sequence                    1300 non-null   object
 10  morph_sequence                  1300 non-null   object
 11  morph_complexity                1300 non-null   float64
```

```
 12  binary_complexity             1300 non-null   int64
 13  binary_complexity_75th_split  1300 non-null   int64
 14  snc_pos_seq                   1300 non-null   object
 15  snc_pos_alt                   1300 non-null   object
 16  snc_morph_seq                 1300 non-null   object
 17  snc_morph_alt                 1300 non-null   object
 18  snc_dep_seq                   1300 non-null   object
 19  snc_dep_alt                   1300 non-null   object
 20  snc_morph_complexity_value    1300 non-null   object
dtypes: bool(1), float64(2), int64(2), object(16)
memory usage: 204.5+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 21 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   id                            1000 non-null   object
 1   corpus                        1000 non-null   object
 2   sentence                      1000 non-null   object
 3   token                         998 non-null    object
 4   complexity                    1000 non-null   float64
 5   is_duplicated                 1000 non-null   object
 6   sentence_no_contractions      1000 non-null   object
 7   contraction_expanded          1000 non-null   bool
 8   pos_sequence                  1000 non-null   object
 9   dep_sequence                  1000 non-null   object
 10  morph_sequence                1000 non-null   object
 11  morph_complexity              1000 non-null   float64
 12  binary_complexity             1000 non-null   int64
 13  binary_complexity_75th_split  1000 non-null   int64
 14  snc_pos_seq                   1000 non-null   object
 15  snc_pos_alt                   1000 non-null   object
 16  snc_morph_seq                 1000 non-null   object
 17  snc_morph_alt                 1000 non-null   object
 18  snc_dep_seq                   1000 non-null   object
 19  snc_dep_alt                   1000 non-null   object
 20  snc_morph_complexity_value    1000 non-null   object
dtypes: bool(1), float64(2), int64(2), object(16)
memory usage: 157.4+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 21 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   id                            250 non-null    object
 1   corpus                        250 non-null    object
```

```
 2   sentence                       250 non-null    object
 3   token                          250 non-null    object
 4   complexity                     250 non-null    float64
 5   is_duplicated                  250 non-null    object
 6   sentence_no_contractions       250 non-null    object
 7   contraction_expanded           250 non-null    bool
 8   pos_sequence                   250 non-null    object
 9   dep_sequence                   250 non-null    object
 10  morph_sequence                 250 non-null    object
 11  morph_complexity               250 non-null    float64
 12  binary_complexity              250 non-null    int64
 13  binary_complexity_75th_split   250 non-null    int64
 14  snc_pos_seq                    250 non-null    object
 15  snc_pos_alt                    250 non-null    object
 16  snc_morph_seq                  250 non-null    object
 17  snc_morph_alt                  250 non-null    object
 18  snc_dep_seq                    250 non-null    object
 19  snc_dep_alt                    250 non-null    object
 20  snc_morph_complexity_value     250 non-null    object
dtypes: bool(1), float64(2), int64(2), object(16)
memory usage: 39.4+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 21 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   id                             1000 non-null   object
 1   corpus                         1000 non-null   object
 2   sentence                       1000 non-null   object
 3   token                          1000 non-null   object
 4   complexity                     1000 non-null   float64
 5   is_duplicated                  1000 non-null   object
 6   sentence_no_contractions       1000 non-null   object
 7   contraction_expanded           1000 non-null   bool
 8   pos_sequence                   1000 non-null   object
 9   dep_sequence                   1000 non-null   object
 10  morph_sequence                 1000 non-null   object
 11  morph_complexity               1000 non-null   float64
 12  binary_complexity              1000 non-null   int64
 13  binary_complexity_75th_split   1000 non-null   int64
 14  snc_pos_seq                    1000 non-null   object
 15  snc_pos_alt                    1000 non-null   object
 16  snc_morph_seq                  1000 non-null   object
 17  snc_morph_alt                  1000 non-null   object
 18  snc_dep_seq                    1000 non-null   object
 19  snc_dep_alt                    1000 non-null   object
 20  snc_morph_complexity_value     1000 non-null   object
```

```
dtypes: bool(1), float64(2), int64(2), object(16)
memory usage: 157.4+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 21 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   id                             250 non-null    object
 1   corpus                         250 non-null    object
 2   sentence                       250 non-null    object
 3   token                          250 non-null    object
 4   complexity                     250 non-null    float64
 5   is_duplicated                  250 non-null    object
 6   sentence_no_contractions       250 non-null    object
 7   contraction_expanded           250 non-null    bool
 8   pos_sequence                   250 non-null    object
 9   dep_sequence                   250 non-null    object
 10  morph_sequence                 250 non-null    object
 11  morph_complexity               250 non-null    float64
 12  binary_complexity              250 non-null    int64
 13  binary_complexity_75th_split   250 non-null    int64
 14  snc_pos_seq                    250 non-null    object
 15  snc_pos_alt                    250 non-null    object
 16  snc_morph_seq                  250 non-null    object
 17  snc_morph_alt                  250 non-null    object
 18  snc_dep_seq                    250 non-null    object
 19  snc_dep_alt                    250 non-null    object
 20  snc_morph_complexity_value     250 non-null    object
dtypes: bool(1), float64(2), int64(2), object(16)
memory usage: 39.4+ 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 is_duplicated
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  3IQ9O0AYW6ZPOAQ7VNRXLNM4D1DITZ    biomed  The development of sexually
dimorphic reproduc…           organs    0.250000            {}  The
development of sexually dimorphic reproduc…                  False  [DET, NOUN,
ADP, ADV, ADJ, ADJ, NOUN, AUX, DET…  [det, nsubj, prep, advmod, amod, amod,
pobj, R…  [(Definite=Def, PronType=Art), (Number=Sing), …          1.200000
0                            0  The development of sexually dimorphic
reproduc…  The [DET] development [NOUN] of [ADP] sexually…  The development
of sexually dimorphic reproduc…  The [(Definite=Def|PronType=Art)] development
…  The development of sexually dimorphic reproduc…  The [det] development
[nsubj] of [prep] sexual…  The development of sexually dimorphic reproduc…
1  3PA41K45VN4U7YG4VFEGPOVYAII7PP    biomed  We find that the majority of the
olfactory rec…             usage    0.382353            {}  We find that the
majority of the olfactory rec…                  False  [PRON, VERB, SCONJ, DET,
NOUN, ADP, DET, ADJ, …  [nsubj, ROOT, mark, det, nsubjpass, prep, det,…
[(Case=Nom, Number=Plur, Person=1, PronType=Pr…          1.142857
1                            1  We find that the majority of the olfactory
rec…  We [PRON] find [VERB] that [SCONJ] the [DET] m…  We find that the
majority of the olfactory rec…  We
[(Case=Nom|Number=Plur|Person=1|PronType=Pr…  We find that the majority of the
olfactory rec…  We [nsubj] find [ROOT] that [mark] the [det] m…  We find
that the majority of the olfactory rec…
2  36818Z1KV3D5JB9F4KTTMCUN6U7A3I     bible  His lord was angry, and delivered
him to the t…       tormentors    0.328947            {}  His lord was angry,
and delivered him to the t…                 False  [PRON, NOUN, AUX, ADJ,
PUNCT, CCONJ, VERB, PRO…  [poss, nsubj, ROOT, acomp, punct, cc, conj, do…
[(Gender=Masc, Number=Sing, Person=3, Poss=Yes…          1.956522
1                            0  His lord was angry, and delivered him to the
t…  His [PRON] lord [NOUN] was [AUX] angry, [ADJ] …  His lord was angry, and
delivered him to the t…  His [(Gender=Masc|Number=Sing|Person=3|Poss=Ye…
His lord was angry, and delivered him to the t…  His [poss] lord [nsubj] was
[ROOT] angry, [aco…  His lord was angry, and delivered him to the t…
3  3VJ4PFXFJ37PI5MYJ4PU9LKNJ9SUAF  europarl  The Taiwanese Government has
informed the Coun…  representations    0.315789            {}  The Taiwanese
Government has informed the Coun…                  False  [DET, ADJ, PROPN,
AUX, VERB, DET, PROPN, PUNCT…  [det, amod, nsubj, aux, ROOT, det, dobj,
punct…  [(Definite=Def, PronType=Art), (Degree=Pos), (…          1.432432
1                            0  The Taiwanese Government has informed the
Coun…  The [DET] Taiwanese [ADJ] Government [PROPN] h…  The Taiwanese
Government has informed the Coun…  The [(Definite=Def|PronType=Art)] Taiwanese
[(…  The Taiwanese Government has informed the Coun…  The [det] Taiwanese
[amod] Government [nsubj] …  The Taiwanese Government has informed the Coun…
4  37AQKJ12TXOFXO6IPZQ1ZUODOJPTTP  europarl  However, I too want to thank
everyone who took…         relation    0.267857            {}  However, I too
want to thank everyone who took…                 False  [ADV, PUNCT, PRON,
ADV, VERB, PART, VERB, PRON…  [advmod, punct, nsubj, advmod, ROOT, aux,
xcom…  [(), (PunctType=Comm), (Case=Nom, Number=Sing,…          1.156250
0                            0  However, I too want to thank everyone who
```

```
took…  However, [ADV] I [PUNCT] too [PRON] want [ADV]…  However, I too want
to thank everyone who took…  However, [()] I [(PunctType=Comm)] too [(Case=…
However, I too want to thank everyone who took…  However, [advmod] I [punct]
too [nsubj] want […  However, I too want to thank everyone who took…
                              id    corpus
sentence                     token  complexity is_duplicated
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  3T2EL38U0MK9MPNAD5X3JSYWH8BQXO    biomed  CA = chronic arthritis; CIA =
collagen-induced…      rheumatoid arthritis    0.600000            {}  CA =
chronic arthritis; CIA = collagen-induced…                 False  [PROPN, ADP,
ADJ, NOUN, PUNCT, PROPN, PUNCT, N…  [nmod, punct, amod, ROOT, punct, nmod,
punct, …  [(Number=Sing), (), (Degree=Pos), (Number=Sing…          0.857143
1                             1  CA = chronic arthritis; CIA = collagen-
induced…  CA [PROPN] = [ADP] chronic [ADJ] arthritis; [N…  CA = chronic
arthritis; CIA = collagen-induced…  CA [(Number=Sing)] = [()] chronic
[(Degree=Pos…  CA = chronic arthritis; CIA = collagen-induced…  CA [nmod] =
[punct] chronic [amod] arthritis; …  CA = chronic arthritis; CIA = collagen-
induced…
1  388CL5C1RJN1927IGW7LZKB8JDSLHQ  europarl  Appointments to parliamentary
committees (vote…  parliamentary committees    0.328947            {}
Appointments to parliamentary committees (vote…                 False  [NOUN,
ADP, ADJ, NOUN, PUNCT, VERB, PUNCT, VER…  [nsubj, prep, amod, pobj, punct,
ccomp, punct,…  [(Number=Plur), (), (Degree=Pos), (Number=Plur…
0.888889                  0                                 0  Appointments to
parliamentary committees (vote…  Appointments [NOUN] to [ADP] parliamentary
[AD…  Appointments to parliamentary committees (vote…  Appointments
[(Number=Plur)] to [()] parliamen…  Appointments to parliamentary committees
(vote…  Appointments [nsubj] to [prep] parliamentary […  Appointments to
parliamentary committees (vote…
2  3A3KKYU7P3H3CAKSB7U0000KY58MW4    biomed  The HG9 strain represents a major
epistasis-ba…               mouse model    0.350000            {}  The HG9
strain represents a major epistasis-ba…                 False  [DET, PROPN,
NOUN, VERB, DET, ADJ, NOUN, PUNCT…  [det, compound, nsubj, ROOT, det, amod,
npadvm…  [(Definite=Def, PronType=Art), (Number=Sing), …          1.093750
0                             0  The HG9 strain represents a major epistasis-
ba…  The [DET] HG9 [PROPN] strain [NOUN] represents…  The HG9 strain
represents a major epistasis-ba…  The [(Definite=Def|PronType=Art)] HG9
[(Number…  The HG9 strain represents a major epistasis-ba…  The [det] HG9
[compound] strain [nsubj] repres…  The HG9 strain represents a major
epistasis-ba…
3  3FBEFUUYRK54GUWXNMRRTF67GLFA6U     bible  For there is an annulling of a
foregoing comma…     foregoing commandment    0.638889            {}  For
```

```
there is an annulling of a foregoing comma…                  False  [ADP, PRON,
VERB, DET, NOUN, ADP, DET, NOUN, N…  [prep, expl, ROOT, det, attr, prep, det,
compo…  [(), (), (Mood=Ind, Number=Sing, Person=3, Ten…          1.333333
1                             1  For there is an annulling of a foregoing
comma…  For [ADP] there [PRON] is [VERB] an [DET] annu…  For there is an
annulling of a foregoing comma…  For [()] there [()] is
[(Mood=Ind|Number=Sing|…  For there is an annulling of a foregoing comma…
For [prep] there [expl] is [ROOT] an [det] ann…  For there is an annulling of
a foregoing comma…
4  36QZ6V1589DTI18S04BLULET5D3SU9     bible  Ezra the priest, with certain heads
of fathers…                  first day    0.116667            {}  Ezra the
priest, with certain heads of fathers…                  False  [PROPN, DET,
NOUN, PUNCT, ADP, ADJ, NOUN, ADP,…  [nsubjpass, det, appos, punct, prep, amod,
pob…  [(Number=Sing), (Definite=Def, PronType=Art), …          1.148936
0                             0  Ezra the priest, with certain heads of
fathers…  Ezra [PROPN] the [DET] priest, [NOUN] with [PU…  Ezra the priest,
with certain heads of fathers…  Ezra [(Number=Sing)] the
[(Definite=Def|PronTy…  Ezra the priest, with certain heads of fathers…
Ezra [nsubjpass] the [det] priest, [appos] wit…  Ezra the priest, with certain
heads of fathers…
                              id  corpus
sentence    token  complexity is_duplicated
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  3ZQA3IO31BRYBCP1RZKSZEZVXRG1OZ  biomed  In addition to colorectal neoplams,
these indi…  pigment    0.350000            {}  In addition to colorectal
neoplams, these indi…                  False  [ADP, NOUN, ADP, ADJ, NOUN,
PUNCT, DET, NOUN, …  [prep, pobj, prep, amod, pobj, punct, det, nsu…  [(),
(Number=Sing), (), (Degree=Pos), (Number=…          1.050847
1                             0  In addition to colorectal neoplams, these
indi…  In [ADP] addition [NOUN] to [ADP] colorectal […  In addition to
colorectal neoplams, these indi…  In [()] addition [(Number=Sing)] to [()]
color…  In addition to colorectal neoplams, these indi…  In [prep] addition
[pobj] to [prep] colorectal…  In addition to colorectal neoplams, these
indi…
1  3Z3R5YC0P3N5EJOHUFLCIQ9CX7PTFJ   bible  The Queen of the South will rise up
in the jud…     ends    0.302632            {}  The Queen of the South will
rise up in the jud…                  False  [DET, PROPN, ADP, DET, PROPN, AUX,
VERB, ADP, …  [det, nsubj, prep, det, pobj, aux, ROOT, prt, …
[(Definite=Def, PronType=Art), (Number=Sing), …          1.142857
1                             0  The Queen of the South will rise up in the
jud…  The [DET] Queen [PROPN] of [ADP] the [DET] Sou…  The Queen of the
South will rise up in the jud…  The [(Definite=Def|PronType=Art)] Queen
```

```
[(Numb…  The Queen of the South will rise up in the jud…  The [det] Queen
[nsubj] of [prep] the [det] So…  The Queen of the South will rise up in the
jud…
2  3URJ6VVYUPNF3BMKEH3UXC6Y9BQ4OF  biomed  Since the parental strains differ in
susceptib…    class    0.261905           {}  Since the parental strains
differ in susceptib…                False  [SCONJ, DET, ADJ, NOUN, VERB, ADP,
NOUN, ADP, …  [mark, det, amod, nsubj, advcl, prep, pobj, pr…  [(),
(Definite=Def, PronType=Art), (Degree=Pos…          1.073171
0                            0  Since the parental strains differ in
susceptib…  Since [SCONJ] the [DET] parental [ADJ] strains…  Since the
parental strains differ in susceptib…  Since [()] the
[(Definite=Def|PronType=Art)] p…  Since the parental strains differ in
susceptib…  Since [mark] the [det] parental [amod] strains…  Since the
parental strains differ in susceptib…
3  3MVY4USGB6NO9ADS6NM7BIQIBGKSI1   bible  For the judgment is against you; for
you have …    Tabor    0.633333           {}  For the judgment is against
you; for you have …                 False  [ADP, DET, NOUN, AUX, ADP, PRON,
PUNCT, SCONJ,…  [prep, det, pobj, ccomp, prep, pobj, punct, ma…  [(),
(Definite=Def, PronType=Art), (Number=Sin…          1.347826
1                            1  For the judgment is against you; for you have
…  For [ADP] the [DET] judgment [NOUN] is [AUX] a…  For the judgment is
against you; for you have …  For [()] the [(Definite=Def|PronType=Art)] jud…
For the judgment is against you; for you have …  For [prep] the [det] judgment
[pobj] is [ccomp…  For the judgment is against you; for you have …
4  30U1YOGZGAW71ZX6E9LWKLA5JD8SDZ   bible  having a great and high wall; having
twelve ga…   tribes    0.175000           {}  having a great and high wall;
having twelve ga…                 False  [VERB, DET, ADJ, CCONJ, ADJ, NOUN,
PUNCT, VERB…  [ROOT, det, amod, cc, conj, dobj, punct, conj,…
[(Aspect=Prog, Tense=Pres, VerbForm=Part), (De…          1.236842
0                            0  having a great and high wall; having twelve
ga…  having [VERB] a [DET] great [ADJ] and [CCONJ] …  having a great and
high wall; having twelve ga…  having
[(Aspect=Prog|Tense=Pres|VerbForm=Part)…  having a great and high wall; having
twelve ga…  having [ROOT] a [det] great [amod] and [cc] hi…  having a great
and high wall; having twelve ga…
                              id    corpus
sentence                token  complexity is_duplicated
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  3D17ECOUOEV9PNWF8100BB1K20731T     bible  But some of the itinerant Jews,
exorcists, too…       itinerant Jews    0.600000            {}  But some of
the itinerant Jews, exorcists, too…                 False  [CCONJ, PRON, ADP,
DET, ADJ, PROPN, PUNCT, NOU…  [cc, nsubj, prep, det, amod, pobj, punct,
```

appo… [(ConjType=Cmp), (), (), (Definite=Def, PronTy… 1.365854
1 1 But some of the itinerant Jews, exorcists,
too… But [CCONJ] some [PRON] of [ADP] the [DET] iti… But some of the
itinerant Jews, exorcists, too… But [(ConjType=Cmp)] some [()] of [()] the
[(D… But some of the itinerant Jews, exorcists, too… But [cc] some [nsubj]
of [prep] the [det] itin… But some of the itinerant Jews, exorcists, too…
1 3XBXDSS888JYVS7XL0P726Z273BLXJ europarl The next item is the report by
Esther de Lange… EU legislation 0.285714 {} The next item
is the report by Esther de Lange… False [DET, ADJ, NOUN, AUX,
DET, NOUN, ADP, PROPN, P… [det, amod, nsubj, ROOT, det, attr, prep, comp…
[(Definite=Def, PronType=Art), (Degree=Pos), (… 1.102564
0 0 The next item is the report by Esther de
Lange… The [DET] next [ADJ] item [NOUN] is [AUX] the … The next item is
the report by Esther de Lange… The [(Definite=Def|PronType=Art)] next
[(Degre… The next item is the report by Esther de Lange… The [det] next
[amod] item [nsubj] is [ROOT] t… The next item is the report by Esther de
Lange…
2 3GITHABACYLNIC7L9OKTP89VZONN2N biomed Alternatively, the unusual
transcriptional reg… olfactory receptors 0.725000 {}
Alternatively, the unusual transcriptional reg… False [ADV,
PUNCT, DET, ADJ, ADJ, NOUN, ADP, ADJ, NO… [advmod, punct, det, amod, amod,
nsubj, prep, … [(), (PunctType=Comm), (Definite=Def, PronType…
1.260870 1 1 Alternatively, the
unusual transcriptional reg… Alternatively, [ADV] the [PUNCT] unusual
[DET]… Alternatively, the unusual transcriptional reg… Alternatively, [()]
the [(PunctType=Comm)] unu… Alternatively, the unusual transcriptional reg…
Alternatively, [advmod] the [punct] unusual [d… Alternatively, the unusual
transcriptional reg…
3 31MCUE39BKM6T2MIQKL3IY5Q4Q13G6 biomed Genetic disruption of the Dhcr7
results in neo… neonatal lethality 0.547619 {} Genetic
disruption of the Dhcr7 results in neo… False [ADJ, NOUN,
ADP, DET, PROPN, NOUN, ADP, ADJ, N… [amod, ROOT, prep, det, compound, pobj,
prep, … [(Degree=Pos), (Number=Sing), (), (Definite=De… 0.923077
1 1 Genetic disruption of the Dhcr7 results in
neo… Genetic [ADJ] disruption [NOUN] of [ADP] the […] Genetic disruption of
the Dhcr7 results in neo… Genetic [(Degree=Pos)] disruption [(Number=Sin…
Genetic disruption of the Dhcr7 results in neo… Genetic [amod] disruption
[ROOT] of [prep] the… Genetic disruption of the Dhcr7 results in neo…
4 37PGLWGSJT6QLR0K1ED5KWZ8UO3IKA bible In it you shall not sow, neither
reap that whi… undressed vines 0.525000 {} In it you shall
not sow, neither reap that whi… False [ADP, PRON, PRON, AUX,
PART, VERB, PUNCT, CCON… [prep, pobj, nsubj, aux, neg, ROOT, punct, pre…
[(), (Case=Acc, Gender=Neut, Number=Sing, Pers… 1.500000
1 0 In it you shall not sow, neither reap that
whi… In [ADP] it [PRON] you [PRON] shall [AUX] not … In it you shall not
sow, neither reap that whi… In [()] it [(Case=Acc|Gender=Neut|Number=Sing|…
In it you shall not sow, neither reap that whi… In [prep] it [pobj] you
[nsubj] shall [aux] no… In it you shall not sow, neither reap that whi…

```
                              id    corpus
sentence         token  complexity is_duplicated
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  3ZURAPD288N45ZC8SW12CKQH5QPF1R    biomed  We show that in p150CAF-1-depleted
ES cells, w…  perturbation    0.484375            {}  We show that in
p150CAF-1-depleted ES cells, w…                  False  [PRON, VERB, SCONJ,
ADP, ADV, PUNCT, VERB, NOU…  [nsubj, ROOT, mark, prep, npadvmod, punct, amo…
[(Case=Nom, Number=Plur, Person=1, PronType=Pr…          1.133333
1                             1  We show that in p150CAF-1-depleted ES cells,
w…  We [PRON] show [VERB] that [SCONJ] in [ADP] p1…  We show that in
p150CAF-1-depleted ES cells, w…  We
[(Case=Nom|Number=Plur|Person=1|PronType=Pr…  We show that in
p150CAF-1-depleted ES cells, w…  We [nsubj] show [ROOT] that [mark] in [prep]
p…  We show that in p150CAF-1-depleted ES cells, w…
1  36D1BWBEHN1H0UMLXN5TCTKVUXL2M8    biomed  Lung development is a complex
process that inv…       process    0.250000            {}  Lung development is
a complex process that inv…                 False  [PROPN, NOUN, AUX, DET,
ADJ, NOUN, PRON, VERB,…  [compound, nsubj, ROOT, det, amod, attr, nsubj…
[(Number=Sing), (Number=Sing), (Mood=Ind, Numb…          1.407407
0                             0  Lung development is a complex process that
inv…  Lung [PROPN] development [NOUN] is [AUX] a [DE…  Lung development is a
complex process that inv…  Lung [(Number=Sing)] development [(Number=Sing…
Lung development is a complex process that inv…  Lung [compound] development
[nsubj] is [ROOT] …  Lung development is a complex process that inv…
2  3QX22DUVOOHQXLKNLXP4EYH6RZBVME  europarl  That is why we want to introduce
the role of m…          role    0.050000            {}  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…
3  3HXCEECSQMT70MEB5X2ITZH9OICZYL  europarl  (CS) I would just like to emphasise
that this …     groupings    0.210526            {}  (CS) I would just like to
emphasise that this …                  False  [PUNCT, PROPN, PUNCT, PRON, AUX,
ADV, VERB, PA…  [punct, npadvmod, punct, nsubj, aux, advmod, R…
[(PunctSide=Ini, PunctType=Brck), (Number=Sing…          1.254545
0                             0  (CS) I would just like to emphasise that this
…  (CS) [PUNCT] I [PROPN] would [PUNCT] just [PRO…  (CS) I would just like
to emphasise that this …  (CS) [(PunctSide=Ini|PunctType=Brck)] I [(Numb…
```

```
(CS) I would just like to emphasise that this …  (CS) [punct] I [npadvmod]
would [punct] just […  (CS) I would just like to emphasise that this …
4  3WGCNLZJKF877FYC1Q6COKNWTFRD10  europarl  I am from a border county myself
and I am a re…       process    0.183333            {}  I am from a border
county myself and I am a re…                  False  [PRON, AUX, ADP, DET,
NOUN, NOUN, PRON, CCONJ,…  [nsubj, ROOT, prep, det, compound, pobj, npadv…
[(Case=Nom, Number=Sing, Person=1, PronType=Pr…          1.609756
0                             0  I am from a border county myself and I am a
re…  I [PRON] am [AUX] from [ADP] a [DET] border [N…  I am from a border
county myself and I am a re…  I
[(Case=Nom|Number=Sing|Person=1|PronType=Prs…  I am from a border county
myself and I am a re…  I [nsubj] am [ROOT] from [prep] a [det] border…  I am
from a border county myself and I am a re…
                               id    corpus
sentence                 token  complexity is_duplicated
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  3FK4G712NX0D3OGOBZGLFKW5KGISST     bible  He shall put no oil on it, neither
shall he pu…          sin offering    0.450000            {}  He shall put no
oil on it, neither shall he pu…                 False  [PRON, AUX, VERB, DET,
NOUN, ADP, PRON, PUNCT,…  [nsubj, aux, ROOT, det, dobj, prep, pobj, punc…
[(Case=Nom, Gender=Masc, Number=Sing, Person=3…          1.833333
1                             0  He shall put no oil on it, neither shall he
pu…  He [PRON] shall [AUX] put [VERB] no [DET] oil …  He shall put no oil on
it, neither shall he pu…  He [(Case=Nom|Gender=Masc|Number=Sing|Person=3…
He shall put no oil on it, neither shall he pu…  He [nsubj] shall [aux] put
[ROOT] no [det] oil…  He shall put no oil on it, neither shall he pu…
1  3UQVX1UPFSHKXGFE8IIVEWDIRVC02P    biomed  During the last few years the
Wnt1-Cre transge…         powerful tool    0.305556            {}  During the
last few years the Wnt1-Cre transge…                 False  [ADP, DET, ADJ,
ADJ, NOUN, DET, NUM, PUNCT, NO…  [prep, det, amod, amod, pobj, det, compound,
p…  [(), (Definite=Def, PronType=Art), (Degree=Pos…          1.161290
0                             0  During the last few years the Wnt1-Cre
transge…  During [ADP] the [DET] last [ADJ] few [ADJ] ye…  During the last
few years the Wnt1-Cre transge…  During [()] the [(Definite=Def|PronType=Art)]
…  During the last few years the Wnt1-Cre transge…  During [prep] the [det]
last [amod] few [amod]…  During the last few years the Wnt1-Cre transge…
2  3T2EL38U0MK9MPNAD5X3JSYWH9XXQJ  europarl  The next item is the report by Mrs
Fajon, on b…      external borders    0.343750            {}  The next item is
the report by Mrs Fajon, on b…                  False  [DET, ADJ, NOUN, AUX,
DET, NOUN, ADP, PROPN, P…  [det, amod, nsubj, ROOT, det, attr, prep, comp…
[(Definite=Def, PronType=Art), (Degree=Pos), (…          1.137500
0                             0  The next item is the report by Mrs Fajon, on
```

```
b…  The [DET] next [ADJ] item [NOUN] is [AUX] the …  The next item is the
report by Mrs Fajon, on b…  The [(Definite=Def|PronType=Art)] next [(Degre…
The next item is the report by Mrs Fajon, on b…  The [det] next [amod] item
[nsubj] is [ROOT] t…  The next item is the report by Mrs Fajon, on b…
3  37AQKJ12TXOFXO6IPZQ1ZUODOJMTTM    biomed  The pathogenesis and developmental
relationshi…  pulmonary hypoplasia    0.675000           {}  The pathogenesis
and developmental relationshi…                False  [DET, NOUN, CCONJ, ADJ,
NOUN, ADP, ADJ, NOUN, …  [det, nsubjpass, cc, amod, conj, prep, amod, p…
[(Definite=Def, PronType=Art), (Number=Sing), …          1.400000
1                            1  The pathogenesis and developmental
relationshi…  The [DET] pathogenesis [NOUN] and [CCONJ] deve…  The
pathogenesis and developmental relationshi…  The [(Definite=Def|PronType=Art)]
pathogenesis…  The pathogenesis and developmental relationshi…  The [det]
pathogenesis [nsubjpass] and [cc] de…  The pathogenesis and developmental
relationshi…
4  3NZ1E5QA6Z1DG01BOHHIWKCD28P5B4     bible  Moreover I will make a covenant of
peace with …  everlasting covenant    0.444444           {}  Moreover I will
make a covenant of peace with …                False  [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
1                            0  Moreover I will make a covenant of peace with
…  Moreover [ADV] I [PRON] will [AUX] make [VERB]…  Moreover I will make a
covenant of peace with …  Moreover [()] I [(Case=Nom|Number=Sing|Person=…
Moreover I will make a covenant of peace with …  Moreover [advmod] I [nsubj]
will [aux] make [c…  Moreover I will make a covenant of peace with …
```

[ ]:
```
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 is_duplicated
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   3IQ9O0AYW6ZPOAQ7VNRXLNM4D1DITZ  biomed  The development of sexually
dimorphic reproduc…     organs    0.250000           {}  The development of
sexually dimorphic reproduc…                False  [DET, NOUN, ADP, ADV, ADJ,
ADJ, NOUN, AUX, DET…  [det, nsubj, prep, advmod, amod, amod, pobj, R…
```

[(Definite=Def, PronType=Art), (Number=Sing), … 1.200000
0 0 The development of sexually dimorphic
reproduc… The [DET] development [NOUN] of [ADP] sexually… The development
of sexually dimorphic reproduc… The [(Definite=Def|PronType=Art)] development
… The development of sexually dimorphic reproduc… The [det] development
[nsubj] of [prep] sexual… The development of sexually dimorphic reproduc…
1 3PA41K45VN4U7YG4VFEGPOVYAII7PP biomed We find that the majority of the
olfactory rec… usage 0.382353 {} We find that the majority
of the olfactory rec… False [PRON, VERB, SCONJ, DET, NOUN,
ADP, DET, ADJ, … [nsubj, ROOT, mark, det, nsubjpass, prep, det,…
[(Case=Nom, Number=Plur, Person=1, PronType=Pr… 1.142857
1 1 We find that the majority of the olfactory
rec… We [PRON] find [VERB] that [SCONJ] the [DET] m… We find that the
majority of the olfactory rec… We
[(Case=Nom|Number=Plur|Person=1|PronType=Pr… We find that the majority of the
olfactory rec… We [nsubj] find [ROOT] that [mark] the [det] m… We find
that the majority of the olfactory rec…
7 391FPZIE4CM4SSUCPAZMQ77RW26HUX biomed ADAM22 and ADAM23 share highly
homologous sequ… sequences 0.300000 {} ADAM22 and ADAM23 share
highly homologous sequ… False [PROPN, CCONJ, PRON, VERB, ADV,
ADJ, NOUN, ADP… [nmod, cc, nsubj, ROOT, advmod, amod, dobj, pr…
[(Number=Sing), (ConjType=Cmp), (), (VerbForm=… 1.000000
1 0 ADAM22 and ADAM23 share highly homologous
sequ… ADAM22 [PROPN] and [CCONJ] ADAM23 [PRON] share… ADAM22 and ADAM23
share highly homologous sequ… ADAM22 [(Number=Sing)] and [(ConjType=Cmp)]
AD… ADAM22 and ADAM23 share highly homologous sequ… ADAM22 [nmod] and [cc]
ADAM23 [nsubj] share [R… ADAM22 and ADAM23 share highly homologous sequ…
8 3V7ICJJAZAGVKHXBACY8RS6Z2C7B4I biomed Raising intracellular Ca2+ led to
relocation o… Raising 0.200000 {} Raising intracellular Ca2+
led to relocation o… False [VERB, ADJ, NOUN, CCONJ, VERB,
ADP, NOUN, ADP,… [ROOT, amod, dobj, cc, conj, prep, pobj, prep,…
[(Aspect=Prog, Tense=Pres, VerbForm=Part), (De… 1.200000
0 0 Raising intracellular Ca2+ led to relocation
o… Raising [VERB] intracellular [ADJ] Ca2+ [NOUN]… Raising intracellular
Ca2+ led to relocation o… Raising [(Aspect=Prog|Tense=Pres|VerbForm=Part…
Raising intracellular Ca2+ led to relocation o… Raising [ROOT] intracellular
[amod] Ca2+ [dobj… Raising intracellular Ca2+ led to relocation o…
11 33P2GD6NRNSQPWP0VWVKKKYTUC0KHX biomed The speed congenic strains developed
herein co… obesity 0.297619 {} The speed congenic strains
developed herein co… False [DET, NOUN, ADJ, NOUN, VERB, ADV,
VERB, ADV, V… [det, nmod, amod, nsubj, acl, advmod, ROOT, ad…
[(Definite=Def, PronType=Art), (Number=Sing), … 1.285714
1 0 The speed congenic strains developed herein
co… The [DET] speed [NOUN] congenic [ADJ] strains … The speed congenic
strains developed herein co… The [(Definite=Def|PronType=Art)] speed
[(Numb… The speed congenic strains developed herein co… The [det] speed
[nmod] congenic [amod] strains… The speed congenic strains developed herein
co…

```
                                id  corpus
sentence                  token  complexity is_duplicated
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   3T2EL38UOMK9MPNAD5X3JSYWH8BQXO  biomed  CA = chronic arthritis; CIA =
collagen-induced…  rheumatoid arthritis    0.600000            {}  CA =
chronic arthritis; CIA = collagen-induced…                  False  [PROPN, ADP,
ADJ, NOUN, PUNCT, PROPN, PUNCT, N…  [nmod, punct, amod, ROOT, punct, nmod,
punct, …  [(Number=Sing), (), (Degree=Pos), (Number=Sing…          0.857143
1                              1  CA = chronic arthritis; CIA = collagen-
induced…  CA [PROPN] = [ADP] chronic [ADJ] arthritis; [N…  CA = chronic
arthritis; CIA = collagen-induced…  CA [(Number=Sing)] = [()] chronic
[(Degree=Pos…  CA = chronic arthritis; CIA = collagen-induced…  CA [nmod] =
[punct] chronic [amod] arthritis; …  CA = chronic arthritis; CIA = collagen-
induced…
2   3A3KKYU7P3H3CAKSB7U0000KY58MW4  biomed  The HG9 strain represents a major
epistasis-ba…           mouse model    0.350000            {}  The HG9 strain
represents a major epistasis-ba…                  False  [DET, PROPN, NOUN,
VERB, DET, ADJ, NOUN, PUNCT…  [det, compound, nsubj, ROOT, det, amod,
npadvm…  [(Definite=Def, PronType=Art), (Number=Sing), …          1.093750
0                              0  The HG9 strain represents a major epistasis-
ba…  The [DET] HG9 [PROPN] strain [NOUN] represents…  The HG9 strain
represents a major epistasis-ba…  The [(Definite=Def|PronType=Art)] HG9
[(Number…  The HG9 strain represents a major epistasis-ba…  The [det] HG9
[compound] strain [nsubj] repres…  The HG9 strain represents a major
epistasis-ba…
11  3KQC8JMJGCSKTYHTAQ3L3YHRJQEH3J  biomed  Subsequently, CNS apoptosis was
shown to be an…     placental defects    0.500000            {}  Subsequently,
CNS apoptosis was shown to be an…                 False  [ADV, PUNCT, PROPN,
NOUN, AUX, VERB, PART, AUX…  [advmod, punct, compound, nsubjpass, auxpass, …
[(), (PunctType=Comm), (Number=Sing), (Number=…          1.190476
1                              0  Subsequently, CNS apoptosis was shown to be
an…  Subsequently, [ADV] CNS [PUNCT] apoptosis [PRO…  Subsequently, CNS
apoptosis was shown to be an…  Subsequently, [()] CNS [(PunctType=Comm)]
apop…  Subsequently, CNS apoptosis was shown to be an…  Subsequently,
[advmod] CNS [punct] apoptosis […  Subsequently, CNS apoptosis was shown to be
an…
15  3EHVO81VN5LOJV3ENMP2F52UL611HC  biomed  Females and males inherit (on
average) the sam…          disease risk    0.319444            {}  Females and
males inherit (on average) the sam…                 False  [NOUN, CCONJ, NOUN,
VERB, PUNCT, ADP, ADJ, PUN…  [nsubj, cc, conj, ROOT, punct, prep, amod, pun…
[(Number=Plur), (ConjType=Cmp), (Number=Plur),…          1.250000
0                              0  Females and males inherit (on average) the
```

```
sam…  Females [NOUN] and [CCONJ] males [NOUN] inheri…  Females and males
inherit (on average) the sam…  Females [(Number=Plur)] and [(ConjType=Cmp)]
m…  Females and males inherit (on average) the sam…  Females [nsubj] and
[cc] males [conj] inherit …  Females and males inherit (on average) the sam…
20  30IRMPJWDZJ3EQ33R17EYOOZHLYRK1  biomed  CIA = collagen-induced arthritis;
CII = collag…         collagen type    0.588235            {}  CIA = collagen-
induced arthritis; CII = collag…                 False  [PROPN, PUNCT, NOUN,
PUNCT, VERB, NOUN, PUNCT,…  [nmod, punct, npadvmod, punct, amod, ROOT, pun…
[(Number=Sing), (PunctType=Comm), (Number=Sing…          0.883721
1                             1  CIA = collagen-induced arthritis; CII =
collag…  CIA [PROPN] = [PUNCT] collagen-induced [NOUN] …  CIA = collagen-
induced arthritis; CII = collag…  CIA [(Number=Sing)] = [(PunctType=Comm)]
colla…  CIA = collagen-induced arthritis; CII = collag…  CIA [nmod] =
[punct] collagen-induced [npadvmo…  CIA = collagen-induced arthritis; CII =
collag…
                              id  corpus
sentence        token  complexity is_duplicated
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  3ZQA3IO31BRYBCP1RZKSZEZVXRG1OZ  biomed  In addition to colorectal neoplams,
these indi…      pigment    0.350000            {}  In addition to colorectal
neoplams, these indi…                 False  [ADP, NOUN, ADP, ADJ, NOUN,
PUNCT, DET, NOUN, …  [prep, pobj, prep, amod, pobj, punct, det, nsu…  [(),
(Number=Sing), (), (Degree=Pos), (Number=…          1.050847
1                             0  In addition to colorectal neoplams, these
indi…  In [ADP] addition [NOUN] to [ADP] colorectal […  In addition to
colorectal neoplams, these indi…  In [()] addition [(Number=Sing)] to [()]
color…  In addition to colorectal neoplams, these indi…  In [prep] addition
[pobj] to [prep] colorectal…  In addition to colorectal neoplams, these
indi…
2  3URJ6VVYUPNF3BMKEH3UXC6Y9BQ4OF  biomed  Since the parental strains differ in
susceptib…        class    0.261905            {}  Since the parental strains
differ in susceptib…                 False  [SCONJ, DET, ADJ, NOUN, VERB, ADP,
NOUN, ADP, …  [mark, det, amod, nsubj, advcl, prep, pobj, pr…  [(),
(Definite=Def, PronType=Art), (Degree=Pos…          1.073171
0                             0  Since the parental strains differ in
susceptib…  Since [SCONJ] the [DET] parental [ADJ] strains…  Since the
parental strains differ in susceptib…  Since [()] the
[(Definite=Def|PronType=Art)] p…  Since the parental strains differ in
susceptib…  Since [mark] the [det] parental [amod] strains…  Since the
parental strains differ in susceptib…
6  3J6BHNXOU9SIZSBBYUQXP4VPGNBNK9  biomed  In this study, the lack of evidence
of abnorma…  development    0.194444            {}  In this study, the lack of
```

```
evidence of abnorma…                  False  [ADP, DET, NOUN, PUNCT, DET, NOUN,
ADP, NOUN, …  [prep, det, pobj, punct, det, nsubjpass, prep,…  [(),
(Number=Sing, PronType=Dem), (Number=Sing…          1.313725
0                             0  In this study, the lack of evidence of
abnorma…  In [ADP] this [DET] study, [NOUN] the [PUNCT] …  In this study,
the lack of evidence of abnorma…  In [()] this [(Number=Sing|PronType=Dem)]
stud…  In this study, the lack of evidence of abnorma…  In [prep] this [det]
study, [pobj] the [punct]…  In this study, the lack of evidence of abnorma…
7  3I6NF2WGIGW97H9M439WXV3AILZG5E  biomed  TRIP13 was originally discovered to
be an inte…     receptor    0.375000            {}  TRIP13 was originally
discovered to be an inte…                 False  [NOUN, AUX, ADV, VERB, PART,
AUX, DET, NOUN, A…  [nsubjpass, auxpass, advmod, ROOT, aux, xcomp,…
[(Number=Sing), (Mood=Ind, Number=Sing, Person…          1.242424
1                             0  TRIP13 was originally discovered to be an
inte…  TRIP13 [NOUN] was [AUX] originally [ADV] disco…  TRIP13 was
originally discovered to be an inte…  TRIP13 [(Number=Sing)] was
[(Mood=Ind|Number=S…  TRIP13 was originally discovered to be an inte…
TRIP13 [nsubjpass] was [auxpass] originally [a…  TRIP13 was originally
discovered to be an inte…
9  306996CF6WKESI0SNUF6TUZWQS2B1A  biomed  However, the process appears to be
patterned a…    effectors    0.500000            {}  However, the process
appears to be patterned a…                 False  [ADV, PUNCT, DET, NOUN,
VERB, PART, AUX, VERB,…  [advmod, punct, det, nsubj, ROOT, aux, auxpass…
[(), (PunctType=Comm), (Definite=Def, PronType…          1.256410
1                             1  However, the process appears to be patterned
a…  However, [ADV] the [PUNCT] process [DET] appea…  However, the process
appears to be patterned a…  However, [()] the [(PunctType=Comm)] process […
However, the process appears to be patterned a…  However, [advmod] the [punct]
process [det] ap…  However, the process appears to be patterned a…
                              id  corpus
sentence                token  complexity is_duplicated
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
2   3GITHABACYLNIC7L9OKTP89VZONN2N  biomed  Alternatively, the unusual
transcriptional reg…  olfactory receptors    0.725000            {}
Alternatively, the unusual transcriptional reg…                False  [ADV,
PUNCT, DET, ADJ, ADJ, NOUN, ADP, ADJ, NO…  [advmod, punct, det, amod, amod,
nsubj, prep, …  [(), (PunctType=Comm), (Definite=Def, PronType…
1.260870                  1                                1  Alternatively, the
unusual transcriptional reg…  Alternatively, [ADV] the [PUNCT] unusual
[DET]…  Alternatively, the unusual transcriptional reg…  Alternatively, [()]
the [(PunctType=Comm)] unu…  Alternatively, the unusual transcriptional reg…
Alternatively, [advmod] the [punct] unusual [d…  Alternatively, the unusual
```

```
transcriptional reg…
3   31MCUE39BKM6T2MIQKL3IY5Q4Q13G6  biomed  Genetic disruption of the Dhcr7
results in neo…   neonatal lethality    0.547619            {}  Genetic
disruption of the Dhcr7 results in neo…                 False  [ADJ, NOUN,
ADP, DET, PROPN, NOUN, ADP, ADJ, N…  [amod, ROOT, prep, det, compound, pobj,
prep, …  [(Degree=Pos), (Number=Sing), (), (Definite=De…          0.923077
1                             1  Genetic disruption of the Dhcr7 results in
neo…  Genetic [ADJ] disruption [NOUN] of [ADP] the […  Genetic disruption of
the Dhcr7 results in neo…  Genetic [(Degree=Pos)] disruption [(Number=Sin…
Genetic disruption of the Dhcr7 results in neo…  Genetic [amod] disruption
[ROOT] of [prep] the…  Genetic disruption of the Dhcr7 results in neo…
7   3CRWSLD91K4V71BQKL3QJ6NYV39MOS  biomed  If it is true that m-calpain is
essential for …       cell viability    0.406250            {}  If it is true
that m-calpain is essential for …                 False  [SCONJ, PRON, AUX,
ADJ, SCONJ, NOUN, PUNCT, NO…  [mark, nsubj, advcl, acomp, mark, compound,
pu…  [(), (Case=Nom, Gender=Neut, Number=Sing, Pers…          1.156863
0                             0  If it is true that m-calpain is essential for
…  If [SCONJ] it [PRON] is [AUX] true [ADJ] that …  If it is true that
m-calpain is essential for …  If [()] it
[(Case=Nom|Gender=Neut|Number=Sing|…  If it is true that m-calpain is
essential for …  If [mark] it [nsubj] is [advcl] true [acomp] t…  If it is
true that m-calpain is essential for …
12  3Y4OHMYLL1I1EIURUEH8TTVLKTGUXW  biomed  Mutations in either Gdf5 or the
closely relate…   specific locations    0.289474            {}  Mutations in
either Gdf5 or the closely relate…                 False  [NOUN, ADP, DET,
NOUN, CCONJ, DET, ADV, VERB, …  [nsubj, prep, det, pobj, cc, det, advmod,
amod…  [(Number=Plur), (), (), (Number=Plur), (ConjTy…          1.065217
0                             0  Mutations in either Gdf5 or the closely
relate…  Mutations [NOUN] in [ADP] either [DET] Gdf5 [N…  Mutations in
either Gdf5 or the closely relate…  Mutations [(Number=Plur)] in [()] either
[()] …  Mutations in either Gdf5 or the closely relate…  Mutations [nsubj]
in [prep] either [det] Gdf5 …  Mutations in either Gdf5 or the closely
relate…
23  3KWGG5KP6J2UYCENUGUZO6TH6PQCML  biomed  Concerning odor discrimination
itself, cellula…  cellular mechanisms    0.575000            {}  Concerning
odor discrimination itself, cellula…                False  [VERB, NOUN, NOUN,
PRON, PUNCT, ADJ, NOUN, ADP…  [advcl, compound, pobj, appos, punct, amod,
ns…  [(Aspect=Prog, Tense=Pres, VerbForm=Part), (Nu…          1.324324
1                             1  Concerning odor discrimination itself,
cellula…  Concerning [VERB] odor [NOUN] discrimination […  Concerning odor
discrimination itself, cellula…  Concerning
[(Aspect=Prog|Tense=Pres|VerbForm=P…  Concerning odor discrimination itself,
cellula…  Concerning [advcl] odor [compound] discriminat…  Concerning odor
discrimination itself, cellula…
                               id  corpus
sentence         token  complexity is_duplicated
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   3ZURAPD288N45ZC8SW12CKQH5QPF1R  biomed  We show that in p150CAF-1-depleted
ES cells, w…  perturbation    0.484375            {}  We show that in
p150CAF-1-depleted ES cells, w…                  False  [PRON, VERB, SCONJ,
ADP, ADV, PUNCT, VERB, NOU…  [nsubj, ROOT, mark, prep, npadvmod, punct, amo…
[(Case=Nom, Number=Plur, Person=1, PronType=Pr…          1.133333
1                             1  We show that in p150CAF-1-depleted ES cells,
w…  We [PRON] show [VERB] that [SCONJ] in [ADP] p1…  We show that in
p150CAF-1-depleted ES cells, w…  We
[(Case=Nom|Number=Plur|Person=1|PronType=Pr…  We show that in
p150CAF-1-depleted ES cells, w…  We [nsubj] show [ROOT] that [mark] in [prep]
p…  We show that in p150CAF-1-depleted ES cells, w…
1   36D1BWBEHN1HOUMLXN5TCTKVUXL2M8  biomed  Lung development is a complex
process that inv…       process    0.250000            {}  Lung development is
a complex process that inv…                 False  [PROPN, NOUN, AUX, DET,
ADJ, NOUN, PRON, VERB,…  [compound, nsubj, ROOT, det, amod, attr, nsubj…
[(Number=Sing), (Number=Sing), (Mood=Ind, Numb…          1.407407
0                             0  Lung development is a complex process that
inv…  Lung [PROPN] development [NOUN] is [AUX] a [DE…  Lung development is a
complex process that inv…  Lung [(Number=Sing)] development [(Number=Sing…
Lung development is a complex process that inv…  Lung [compound] development
[nsubj] is [ROOT] …  Lung development is a complex process that inv…
6   3KWGG5KP6J2UYCENUGUZO6TH6OSCML  biomed  The pre-publication history for this
paper can…       history    0.170455            {}  The pre-publication
history for this paper can…                 False  [DET, ADJ, NOUN, ADJ, NOUN,
ADP, DET, NOUN, AU…  [det, amod, amod, amod, nsubjpass, prep, det, …
[(Definite=Def, PronType=Art), (Degree=Pos), (…          1.153846
0                             0  The pre-publication history for this paper
can…  The [DET] pre-publication [ADJ] history [NOUN]…  The pre-publication
history for this paper can…  The [(Definite=Def|PronType=Art)] pre-publicat…
The pre-publication history for this paper can…  The [det] pre-publication
[amod] history [amod…  The pre-publication history for this paper can…
9   3MXX6RQ9EV5XOBYLTHG9MCBOJGRP4M  biomed  Future direct comparison of the two
mouse line…      interest    0.279412            {}  Future direct comparison
of the two mouse line…                  False  [ADJ, ADJ, NOUN, ADP, DET, NUM,
NOUN, NOUN, AD…  [amod, amod, nsubj, prep, det, nummod, compoun…
[(Degree=Pos), (Degree=Pos), (Number=Sing), ()…          0.882353
0                             0  Future direct comparison of the two mouse
line…  Future [ADJ] direct [ADJ] comparison [NOUN] of…  Future direct
comparison of the two mouse line…  Future [(Degree=Pos)] direct [(Degree=Pos)]
co…  Future direct comparison of the two mouse line…  Future [amod] direct
[amod] comparison [nsubj]…  Future direct comparison of the two mouse line…
10  3TKSOBLOHLGF5GIKPR8VZ6C5C4OBBO  biomed  Higher expression levels could be
due to incre…    transcript    0.357143            {}  Higher expression
```

```
levels could be due to incre…                  False  [ADJ, NOUN, NOUN, AUX,
AUX, ADJ, ADP, VERB, NO…  [amod, compound, nsubj, aux, ROOT, acomp, pcom…
[(Degree=Cmp), (Number=Sing), (Number=Plur), (…          1.200000
1                             0  Higher expression levels could be due to
incre…  Higher [ADJ] expression [NOUN] levels [NOUN] c…  Higher expression
levels could be due to incre…  Higher [(Degree=Cmp)] expression
[(Number=Sing…  Higher expression levels could be due to incre…  Higher
[amod] expression [compound] levels [ns…  Higher expression levels could be
due to incre…
                                id  corpus
sentence                   token  complexity is_duplicated
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
1   3UQVX1UPFSHKXGFE8IIVEWDIRVC02P  biomed  During the last few years the
Wnt1-Cre transge…          powerful tool    0.305556            {}  During the
last few years the Wnt1-Cre transge…                  False  [ADP, DET, ADJ,
ADJ, NOUN, DET, NUM, PUNCT, NO…  [prep, det, amod, amod, pobj, det, compound,
p…  [(), (Definite=Def, PronType=Art), (Degree=Pos…          1.161290
0                             0  During the last few years the Wnt1-Cre
transge…  During [ADP] the [DET] last [ADJ] few [ADJ] ye…  During the last
few years the Wnt1-Cre transge…  During [()] the [(Definite=Def|PronType=Art)]
…  During the last few years the Wnt1-Cre transge…  During [prep] the [det]
last [amod] few [amod]…  During the last few years the Wnt1-Cre transge…
3   37AQKJ12TXOFXO6IPZQ1ZUODOJMTTM  biomed  The pathogenesis and developmental
relationshi…   pulmonary hypoplasia    0.675000            {}  The
pathogenesis and developmental relationshi…                  False  [DET, NOUN,
CCONJ, ADJ, NOUN, ADP, ADJ, NOUN, …  [det, nsubjpass, cc, amod, conj, prep,
amod, p…  [(Definite=Def, PronType=Art), (Number=Sing), …          1.400000
1                             1  The pathogenesis and developmental
relationshi…  The [DET] pathogenesis [NOUN] and [CCONJ] deve…  The
pathogenesis and developmental relationshi…  The [(Definite=Def|PronType=Art)]
pathogenesis…  The pathogenesis and developmental relationshi…  The [det]
pathogenesis [nsubjpass] and [cc] de…  The pathogenesis and developmental
relationshi…
8   301KG0KX9CLV8GLA6QPGKOCZDCE2HT  biomed  Early migratory CNCCs have been
shown to retai…  instructional signals    0.361111            {}  Early
migratory CNCCs have been shown to retai…                 False  [ADJ, ADJ,
NOUN, AUX, AUX, VERB, PART, VERB, D…  [amod, compound, nsubjpass, aux,
auxpass, ROOT…  [(Degree=Pos), (Degree=Pos), (Number=Plur), (M…
1.272727                  0                             0  Early migratory CNCCs
have been shown to retai…  Early [ADJ] migratory [ADJ] CNCCs [NOUN] have …
Early migratory CNCCs have been shown to retai…  Early [(Degree=Pos)]
migratory [(Degree=Pos)] …  Early migratory CNCCs have been shown to retai…
```

```
Early [amod] migratory [compound] CNCCs [nsubj…  Early migratory CNCCs have
been shown to retai…
11  3SZYX62S5G0QEOYLB052RIQHJJ975X  biomed  In order to better characterize the
defects in…    chromosomal regions    0.500000            {}  In order to
better characterize the defects in…                 False  [ADP, NOUN, PART,
ADV, VERB, DET, NOUN, ADP, N…  [prep, pobj, aux, advmod, acl, det, dobj,
prep…  [(), (Number=Sing), (), (Degree=Cmp), (VerbFor…          1.000000
1                             0  In order to better characterize the defects
in…  In [ADP] order [NOUN] to [PART] better [ADV] c…  In order to better
characterize the defects in…  In [()] order [(Number=Sing)] to [()] better
[…  In order to better characterize the defects in…  In [prep] order [pobj]
to [aux] better [advmod…  In order to better characterize the defects in…
16  3GKAWYFRAPTAO7HEMSH2PG5UWUDPDU  biomed  In support of this, the expression
of the huma…        lupus nephritis    0.661765            {}  In support of
this, the expression of the huma…                 False  [ADP, NOUN, ADP,
PRON, PUNCT, DET, NOUN, ADP, …  [prep, pobj, prep, pobj, punct, det,
nsubjpass…  [(), (Number=Sing), (), (Number=Sing, PronType…
1.112903                  1                             1  In support of this,
the expression of the huma…  In [ADP] support [NOUN] of [ADP] this, [PRON] …
In support of this, the expression of the huma…  In [()] support
[(Number=Sing)] of [()] this, …  In support of this, the expression of the
huma…  In [prep] support [pobj] of [prep] this, [pobj…  In support of this,
the expression of the huma…
```

[ ]:
```
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'] == 'europarl'].head())
    else:
        pass
```

```
                              id    corpus
sentence            token  complexity is_duplicated
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
3   3VJ4PFXFJ37PI5MYJ4PU9LKNJ9SUAF  europarl  The Taiwanese Government has
informed the Coun…  representations    0.315789            {}  The Taiwanese
Government has informed the Coun…                 False  [DET, ADJ, PROPN,
AUX, VERB, DET, PROPN, PUNCT…  [det, amod, nsubj, aux, ROOT, det, dobj,
punct…  [(Definite=Def, PronType=Art), (Degree=Pos), (…          1.432432
1                             0  The Taiwanese Government has informed the
```

```
Coun…  The [DET] Taiwanese [ADJ] Government [PROPN] h…  The Taiwanese
Government has informed the Coun…  The [(Definite=Def|PronType=Art)] Taiwanese
[(…  The Taiwanese Government has informed the Coun…  The [det] Taiwanese
[amod] Government [nsubj] …  The Taiwanese Government has informed the Coun…
4   37AQKJ12TXOFXO6IPZQ1ZUODOJPTTP  europarl  However, I too want to thank
everyone who took…         relation    0.267857           {}  However, I too
want to thank everyone who took…                 False  [ADV, PUNCT, PRON,
ADV, VERB, PART, VERB, PRON…  [advmod, punct, nsubj, advmod, ROOT, aux,
xcom…  [(), (PunctType=Comm), (Case=Nom, Number=Sing,…          1.156250
0                            0  However, I too want to thank everyone who
took…  However, [ADV] I [PUNCT] too [PRON] want [ADV]…  However, I too want
to thank everyone who took…  However, [()] I [(PunctType=Comm)] too [(Case=…
However, I too want to thank everyone who took…  However, [advmod] I [punct]
too [nsubj] want […  However, I too want to thank everyone who took…
5   3PGQRAZXO2KAZASXA58AX6K615VSYW  europarl  Mr President, the subject of this
debate is su…     Johannesburg    0.500000           {}  Mr President, the
subject of this debate is su…                False  [PROPN, PROPN, PUNCT,
DET, NOUN, ADP, DET, NOU…  [compound, nsubj, punct, det, appos, prep, det…
[(Number=Sing), (Number=Sing), (PunctType=Comm…          1.177778
1                            1  Mr President, the subject of this debate is
su…  Mr [PROPN] President, [PROPN] the [PUNCT] subj…  Mr President, the
subject of this debate is su…  Mr [(Number=Sing)] President, [(Number=Sing)]
…  Mr President, the subject of this debate is su…  Mr [compound] President,
[nsubj] the [punct] s…  Mr President, the subject of this debate is su…
6   3Y3CZJSZ9KT0W7I0KE38WZHHKV75RY  europarl  However, some Council working
parties are alre…           trends    0.203125           {}  However, some
Council working parties are alre…                  False  [ADV, PUNCT, DET,
PROPN, VERB, NOUN, AUX, ADV,…  [advmod, punct, det, nmod, amod, nsubj, aux,
a…  [(), (PunctType=Comm), (), (Number=Sing), (Asp…          1.106383
0                            0  However, some Council working parties are
alre…  However, [ADV] some [PUNCT] Council [DET] work…  However, some
Council working parties are alre…  However, [()] some [(PunctType=Comm)]
Council …  However, some Council working parties are alre…  However,
[advmod] some [punct] Council [det] w…  However, some Council working parties
are alre…
10  3SSN80MU8CONBMPF00D6N6MNI2MXK6  europarl  In fact, there is a discussion
regarding the i…        inclusion    0.125000           {}  In fact, there is
a discussion regarding the i…                False  [ADP, NOUN, PUNCT, PRON,
VERB, DET, NOUN, VERB…  [prep, pobj, punct, expl, ROOT, det, attr, pre…
[(), (Number=Sing), (PunctType=Comm), (), (Moo…          1.266667
0                            0  In fact, there is a discussion regarding the
i…  In [ADP] fact, [NOUN] there [PUNCT] is [PRON] …  In fact, there is a
discussion regarding the i…  In [()] fact, [(Number=Sing)] there [(PunctTyp…
In fact, there is a discussion regarding the i…  In [prep] fact, [pobj] there
[punct] is [expl]…  In fact, there is a discussion regarding the i…
                              id    corpus
sentence                     token  complexity is_duplicated
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
1  388CL5C1RJN1927IGW7LZKB8JDSLHQ  europarl  Appointments to parliamentary
committees (vote…  parliamentary committees    0.328947            {}
Appointments to parliamentary committees (vote…                  False  [NOUN,
ADP, ADJ, NOUN, PUNCT, VERB, PUNCT, VER…  [nsubj, prep, amod, pobj, punct,
ccomp, punct,…  [(Number=Plur), (), (Degree=Pos), (Number=Plur…
0.888889                  0                              0  Appointments to
parliamentary committees (vote…  Appointments [NOUN] to [ADP] parliamentary
[AD…  Appointments to parliamentary committees (vote…  Appointments
[(Number=Plur)] to [()] parliamen…  Appointments to parliamentary committees
(vote…  Appointments [nsubj] to [prep] parliamentary […  Appointments to
parliamentary committees (vote…
6  374UMBUHN5PYB7473DVBAKO90Y9TCS  europarl  Oral questions and written
declarations (submi…            Oral questions    0.285714            {}  Oral
questions and written declarations (submi…                 False  [ADJ, NOUN,
CCONJ, VERB, NOUN, PUNCT, NOUN, PU…  [amod, nsubj, cc, amod, conj, punct,
appos, pu…  [(Degree=Pos), (Number=Plur), (ConjType=Cmp), …
1.300000                  0                              0  Oral questions and
written declarations (submi…  Oral [ADJ] questions [NOUN] and [CCONJ]
writte…  Oral questions and written declarations (submi…  Oral
[(Degree=Pos)] questions [(Number=Plur)] …  Oral questions and written
declarations (submi…  Oral [amod] questions [nsubj] and [cc] written…  Oral
questions and written declarations (submi…
7  306996CF6WKESI0SNUF6TUZWQUIB1U  europarl  The essential issue, however, is
that China ha…            Kyoto protocol    0.489130            {}  The
essential issue, however, is that China ha…                 False  [DET, ADJ,
NOUN, PUNCT, ADV, PUNCT, AUX, SCONJ…  [det, amod, nsubj, punct, advmod, punct,
ROOT,…  [(Definite=Def, PronType=Art), (Degree=Pos), (…          1.177083
1                            0  The essential issue, however, is that China
ha…  The [DET] essential [ADJ] issue, [NOUN] howeve…  The essential issue,
however, is that China ha…  The [(Definite=Def|PronType=Art)] essential [(…
The essential issue, however, is that China ha…  The [det] essential [amod]
issue, [nsubj] howe…  The essential issue, however, is that China ha…
8  3BAKUKE49HC18PHHJR1WT9408GLR1N  europarl  We are a long way from the unequal
treatment o…         unequal treatment    0.315789            {}  We are a
long way from the unequal treatment o…                 False  [PRON, AUX, DET,
ADJ, NOUN, ADP, DET, ADJ, NOU…  [nsubj, ROOT, det, amod, attr, prep, det,
amod…  [(Case=Nom, Number=Plur, Person=1, PronType=Pr…          1.181818
0                            0  We are a long way from the unequal treatment
o…  We [PRON] are [AUX] a [DET] long [ADJ] way [NO…  We are a long way from
the unequal treatment o…  We [(Case=Nom|Number=Plur|Person=1|PronType=Pr…
We are a long way from the unequal treatment o…  We [nsubj] are [ROOT] a [det]
long [amod] way …  We are a long way from the unequal treatment o…
```

```
9  3SX4X51T809U5021NIDLAFSY1N7AOL  europarl  Financing instrument for
development cooperati…   development cooperation    0.368421            {}
Financing instrument for development cooperati…                 False  [NOUN,
NOUN, ADP, NOUN, NOUN, PUNCT, NOUN, PUNCT]  [compound, ROOT, prep, compound,
pobj, punct, …  [(Number=Sing), (Number=Sing), (), (Number=Sin…
1.125000                  0                              0  Financing instrument
for development cooperati…  Financing [NOUN] instrument [NOUN] for [ADP] d…
Financing instrument for development cooperati…  Financing [(Number=Sing)]
instrument [(Number=…  Financing instrument for development cooperati…
Financing [compound] instrument [ROOT] for [pr…  Financing instrument for
development cooperati…
                                id    corpus
sentence      token  complexity is_duplicated
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
10  3JMNNNO3B14D56GZ1PBGLRMMS0V2WX  europarl  The epilogue to this disaster must
surely be t…   adoption    0.214286            {}  The epilogue to this
disaster must surely be t…                  False  [DET, NOUN, ADP, DET, NOUN,
AUX, ADV, AUX, DET…  [det, nsubj, prep, det, pobj, aux, advmod, ROO…
[(Definite=Def, PronType=Art), (Number=Sing), …          1.037037
0                             0  The epilogue to this disaster must surely be
t…  The [DET] epilogue [NOUN] to [ADP] this [DET] …  The epilogue to this
disaster must surely be t…  The [(Definite=Def|PronType=Art)] epilogue [(N…
The epilogue to this disaster must surely be t…  The [det] epilogue [nsubj] to
[prep] this [det…  The epilogue to this disaster must surely be t…
12  3MZ3TAMYTLNC8VDFRYM2L8LMPIVRIE  europarl  It may also have been forgotten
that, in 1990,…     League    0.294118            {}  It may also have been
forgotten that, in 1990,…                  False  [PRON, AUX, ADV, AUX, AUX,
VERB, SCONJ, PUNCT,…  [nsubjpass, aux, advmod, aux, auxpass, ROOT, m…
[(Gender=Neut, Number=Sing, Person=3, PronType…          1.234043
1                             0  It may also have been forgotten that, in
1990,…  It [PRON] may [AUX] also [ADV] have [AUX] been…  It may also have
been forgotten that, in 1990,…  It
[(Gender=Neut|Number=Sing|Person=3|PronType…  It may also have been forgotten
that, in 1990,…  It [nsubjpass] may [aux] also [advmod] have [a…  It may
also have been forgotten that, in 1990,…
15  3YLTXLH3DF6R0NMG8O0SG1KSOM9PH6  europarl  Addressing this crisis is an
important test fo…       test    0.190476            {}  Addressing this
crisis is an important test fo…                 False  [VERB, DET, NOUN, AUX,
DET, ADJ, NOUN, ADP, DE…  [csubj, det, dobj, ROOT, det, amod, attr, prep…
[(Aspect=Prog, Tense=Pres, VerbForm=Part), (Nu…          1.379310
0                             0  Addressing this crisis is an important test
fo…  Addressing [VERB] this [DET] crisis [NOUN] is …  Addressing this crisis
```

```
is an important test fo…  Addressing [(Aspect=Prog|Tense=Pres|VerbForm=P…
Addressing this crisis is an important test fo…  Addressing [csubj] this [det]
crisis [dobj] is…  Addressing this crisis is an important test fo…
22  36KM3FWE3RCRJHCKEUZQANUQ2B270G  europarl  Please allow me to start with some
general rem…    remarks    0.183333            {}  Please allow me to start
with some general rem…                 False  [INTJ, VERB, PRON, PART, VERB,
ADP, DET, ADJ, …  [intj, ROOT, nsubj, aux, ccomp, prep, det, amo…  [(),
(VerbForm=Inf), (Case=Acc, Number=Sing, P…          0.866667
0                             0  Please allow me to start with some general
rem…  Please [INTJ] allow [VERB] me [PRON] to [PART]…  Please allow me to
start with some general rem…  Please [()] allow [(VerbForm=Inf)] me
[(Case=A…  Please allow me to start with some general rem…  Please [intj]
allow [ROOT] me [nsubj] to [aux]…  Please allow me to start with some general
rem…
25  37SDSEDIN92VQK2LKIVW2S9VIOX18L  europarl  However, it is not in the
possession of inform…  violation    0.315789            {}  However, it is not
in the possession of inform…                False  [ADV, PUNCT, PRON, AUX,
PART, ADP, DET, NOUN, …  [advmod, punct, nsubj, ROOT, neg, prep, det, p…
[(), (PunctType=Comm), (Case=Nom, Gender=Neut,…          1.344828
1                             0  However, it is not in the possession of
inform…  However, [ADV] it [PUNCT] is [PRON] not [AUX] …  However, it is not
in the possession of inform…  However, [()] it [(PunctType=Comm)] is
[(Case=…  However, it is not in the possession of inform…  However, [advmod]
it [punct] is [nsubj] not [R…  However, it is not in the possession of
inform…
                               id    corpus
sentence                     token  complexity is_duplicated
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
1  3XBXDSS888JYVS7XL0P726Z273BLXJ  europarl  The next item is the report by
Esther de Lange…           EU legislation    0.285714            {}  The next
item is the report by Esther de Lange…                 False  [DET, ADJ, NOUN,
AUX, DET, NOUN, ADP, PROPN, P…  [det, amod, nsubj, ROOT, det, attr, prep,
comp…  [(Definite=Def, PronType=Art), (Degree=Pos), (…          1.102564
0                             0  The next item is the report by Esther de
Lange…  The [DET] next [ADJ] item [NOUN] is [AUX] the …  The next item is
the report by Esther de Lange…  The [(Definite=Def|PronType=Art)] next
[(Degre…  The next item is the report by Esther de Lange…  The [det] next
[amod] item [nsubj] is [ROOT] t…  The next item is the report by Esther de
Lange…
5  3FO95NVK5C0129GBWAGGPAR9YGYSR8  europarl  However, it is unfortunate that the
Russian au…      Russian authorities    0.362500            {}  However, it is
unfortunate that the Russian au…                False  [ADV, PUNCT, PRON,
```

```
AUX, ADJ, SCONJ, DET, ADJ, …  [advmod, punct, nsubj, ROOT, acomp, mark,
det,…  [(), (PunctType=Comm), (Case=Nom, Gender=Neut,…          1.230769
0                             0  However, it is unfortunate that the Russian
au…  However, [ADV] it [PUNCT] is [PRON] unfortunat…  However, it is
unfortunate that the Russian au…  However, [()] it [(PunctType=Comm)] is
[(Case=…  However, it is unfortunate that the Russian au…  However, [advmod]
it [punct] is [nsubj] unfort…  However, it is unfortunate that the Russian
au…
6  3FTID4TN8LYNVXX7QWB9LK6B995LYP  europarl  They also allow for easy
compensation for the …        easy compensation    0.232143           {}
They also allow for easy compensation for the …                 False  [PRON,
ADV, VERB, ADP, ADJ, NOUN, ADP, DET, NO…  [nsubj, advmod, ROOT, prep, amod,
pobj, prep, …  [(Case=Nom, Number=Plur, Person=3, PronType=Pr…
1.050000                  0                             0  They also allow for
easy compensation for the …  They [PRON] also [ADV] allow [VERB] for [ADP] …
They also allow for easy compensation for the …  They
[(Case=Nom|Number=Plur|Person=3|PronType=…  They also allow for easy
compensation for the …  They [nsubj] also [advmod] allow [ROOT] for [p…
They also allow for easy compensation for the …
8  30ZKOOGW2W6998V0HGFAYJFQ8CXA1Y  europarl  This is entirely in line with
international co…  international consensus    0.590909           {}  This is
entirely in line with international co…                 False  [PRON, AUX,
ADV, ADP, NOUN, ADP, ADJ, NOUN, AD…  [nsubj, ROOT, advmod, prep, pobj, prep,
amod, …  [(Number=Sing, PronType=Dem), (Mood=Ind, Numbe…          1.076923
1                             1  This is entirely in line with international
co…  This [PRON] is [AUX] entirely [ADV] in [ADP] l…  This is entirely in
line with international co…  This [(Number=Sing|PronType=Dem)] is [(Mood=In…
This is entirely in line with international co…  This [nsubj] is [ROOT]
entirely [advmod] in [p…  This is entirely in line with international co…
9  3X4Q1O9UBHMCMY43GF110OQ80F1O78  europarl  But let me start with the facts of
the inciden…           EU legislation    0.339286           {}  But let me
start with the facts of the inciden…                 False  [CCONJ, VERB,
PRON, VERB, ADP, DET, NOUN, ADP,…  [cc, ccomp, nsubj, ccomp, prep, det, pobj,
pre…  [(ConjType=Cmp), (VerbForm=Inf), (Case=Acc, Nu…          1.203390
0                             0  But let me start with the facts of the
inciden…  But [CCONJ] let [VERB] me [PRON] start [VERB] …  But let me start
with the facts of the inciden…  But [(ConjType=Cmp)] let [(VerbForm=Inf)] me
[…  But let me start with the facts of the inciden…  But [cc] let [ccomp] me
[nsubj] start [ccomp] …  But let me start with the facts of the inciden…
                               id    corpus
sentence      token  complexity is_duplicated
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
```

2 3QX22DUVOOHQXLKNLXP4EYH6RZBVME europarl That is why we want to introduce the role of m… role 0.050000 {} 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…
3 3HXCEECSQMT70MEB5X2ITZH9OICZYL europarl (CS) I would just like to emphasise that this … groupings 0.210526 {} (CS) I would just like to emphasise that this … False [PUNCT, PROPN, PUNCT, PRON, AUX, ADV, VERB, PA… [punct, npadvmod, punct, nsubj, aux, advmod, R… [(PunctSide=Ini, PunctType=Brck), (Number=Sing… 1.254545 0 0 (CS) I would just like to emphasise that this … (CS) [PUNCT] I [PROPN] would [PUNCT] just [PRO… (CS) I would just like to emphasise that this … (CS) [(PunctSide=Ini|PunctType=Brck)] I [(Numb… (CS) I would just like to emphasise that this … (CS) [punct] I [npadvmod] would [punct] just [… (CS) I would just like to emphasise that this …
4 3WGCNLZJKF877FYC1Q6COKNWTFRD10 europarl I am from a border county myself and I am a re… process 0.183333 {} I am from a border county myself and I am a re… False [PRON, AUX, ADP, DET, NOUN, NOUN, PRON, CCONJ,… [nsubj, ROOT, prep, det, compound, pobj, npadv… [(Case=Nom, Number=Sing, Person=1, PronType=Pr… 1.609756 0 0 I am from a border county myself and I am a re… I [PRON] am [AUX] from [ADP] a [DET] border [N… I am from a border county myself and I am a re… I [(Case=Nom|Number=Sing|Person=1|PronType=Prs… I am from a border county myself and I am a re… I [nsubj] am [ROOT] from [prep] a [det] border… I am from a border county myself and I am a re…
5 3BDORL6HKKDLVSRFZOQGA5NRCCTCRD europarl Situation in Darfur (vote) Situation 0.211538 {} Situation in Darfur (vote) False [NOUN, ADP, PROPN, PUNCT, NOUN, PUNCT] [ROOT, prep, pobj, punct, appos, punct] [(Number=Sing), (), (Number=Sing), (PunctSide=… 1.166667 0 0 Situation in Darfur (vote) [NOUN, ADP, PROPN, … Situation [NOUN] in [ADP] Darfur [PROPN] (vote… Situation in Darfur (vote) [(Number=Sing), (),… Situation [(Number=Sing)] in [()] Darfur [(Num… Situation in Darfur (vote) [ROOT, prep, pobj, … Situation [ROOT] in [prep] Darfur [pobj] (vote… Situation in Darfur (vote) 1.1666666666666667
7 39N6W9XWRDN795J6F5ET8S13DSIYGV europarl They offer some comfort to footwear producers … account 0.216667 {} They offer some comfort to footwear producers … False [PRON, VERB, DET, NOUN, PART, VERB, NOUN, ADP,… [nsubj, ROOT, det, dobj, aux, relcl, dobj, pre… [(Case=Nom, Number=Plur, Person=3, PronType=Pr… 1.065217 0 0 They offer some comfort to footwear producers … They [PRON] offer [VERB] some [DET] comfort [N… They offer some comfort to footwear producers … They [(Case=Nom|Number=Plur|Person=3|PronType=…

```
They offer some comfort to footwear producers …  They [nsubj] offer [ROOT]
some [det] comfort […  They offer some comfort to footwear producers …
                                id    corpus
sentence                  token  complexity is_duplicated
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
2   3T2EL38U0MK9MPNAD5X3JSYWH9XXQJ  europarl  The next item is the report by Mrs
Fajon, on b…       external borders    0.343750            {}  The next item
is the report by Mrs Fajon, on b…                  False  [DET, ADJ, NOUN, AUX,
DET, NOUN, ADP, PROPN, P…  [det, amod, nsubj, ROOT, det, attr, prep, comp…
[(Definite=Def, PronType=Art), (Degree=Pos), (…          1.137500
0                             0  The next item is the report by Mrs Fajon, on
b…  The [DET] next [ADJ] item [NOUN] is [AUX] the …  The next item is the
report by Mrs Fajon, on b…  The [(Definite=Def|PronType=Art)] next [(Degre…
The next item is the report by Mrs Fajon, on b…  The [det] next [amod] item
[nsubj] is [ROOT] t…  The next item is the report by Mrs Fajon, on b…
6   3MD8CKRQZZN836XL9G72X90M2XLJRJ  europarl  What plans does the Commission
have to introdu…          eco labelling    0.355263            {}  What plans
does the Commission have to introdu…                 False  [PRON, VERB, AUX,
DET, PROPN, VERB, PART, VERB…  [nsubj, csubj, aux, det, nsubj, ROOT, aux,
xco…  [(), (Number=Sing, Person=3, Tense=Pres, VerbF…          1.411765
0                             0  What plans does the Commission have to
introdu…  What [PRON] plans [VERB] does [AUX] the [DET] …  What plans does
the Commission have to introdu…  What [()] plans
[(Number=Sing|Person=3|Tense=P…  What plans does the Commission have to
introdu…  What [nsubj] plans [csubj] does [aux] the [det…  What plans does
the Commission have to introdu…
9   338GLSUI43B4ZJB25FGM8LDQVSNFS0  europarl  I should therefore like to
congratulate the Co…  biometric identifiers    0.515625            {}  I
should therefore like to congratulate the Co…                 False  [PRON,
AUX, ADV, VERB, PART, VERB, DET, PROPN,…  [nsubj, aux, advmod, ROOT, aux,
xcomp, det, do…  [(Case=Nom, Number=Sing, Person=1, PronType=Pr…
1.269231                  1                             0  I should therefore
like to congratulate the Co…  I [PRON] should [AUX] therefore [ADV] like
[VE…  I should therefore like to congratulate the Co…  I
[(Case=Nom|Number=Sing|Person=1|PronType=Prs…  I should therefore like to
congratulate the Co…  I [nsubj] should [aux] therefore [advmod] like…  I
should therefore like to congratulate the Co…
12  3BO3NEOQM0HK9ERYPN0GQIWCPE8IAY  europarl  (ES) Mr President, before moving
on to the fin…       profound sadness    0.390625            {}  (ES) Mr
President, before moving on to the fin…                 False  [PUNCT, PROPN,
PUNCT, PROPN, PROPN, PUNCT, ADP…  [punct, nmod, punct, compound, nsubj, punct,
p…  [(PunctSide=Ini, PunctType=Brck), (Number=Sing…          1.151515
```

```
0                              0  (ES) Mr President, before moving on to the
fin…  (ES) [PUNCT] Mr [PROPN] President, [PUNCT] bef…  (ES) Mr President,
before moving on to the fin…  (ES) [(PunctSide=Ini|PunctType=Brck)] Mr
[(Num…  (ES) Mr President, before moving on to the fin…  (ES) [punct] Mr
[nmod] President, [punct] befo…  (ES) Mr President, before moving on to the
fin…
13  322ZSN9Z5GKVG3RSAYPTRMCL8LST4F  europarl  Agriculture as a strategic sector
in the conte…       strategic sector    0.544118            {}  Agriculture as
a strategic sector in the conte…                  False  [NOUN, ADP, DET, ADJ,
NOUN, ADP, DET, NOUN, AD…  [ROOT, prep, det, amod, pobj, prep, det, pobj,…
[(Number=Sing), (), (Definite=Ind, PronType=Ar…          1.066667
1                              1  Agriculture as a strategic sector in the
conte…  Agriculture [NOUN] as [ADP] a [DET] strategic …  Agriculture as a
strategic sector in the conte…  Agriculture [(Number=Sing)] as [()] a
[(Defini…  Agriculture as a strategic sector in the conte…  Agriculture
[ROOT] as [prep] a [det] strategic…  Agriculture as a strategic sector in the
conte…
```

[ ]:
```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})
```

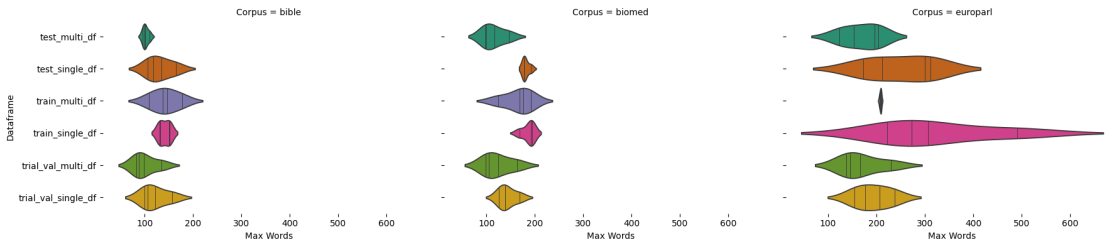
```
            words = tokenizer.tokenize(sentence)
            word_count = len(words)
            char_count = len(sentence)
            avg_word_len = np.mean([len(w) for w in words]) if word_count > 0␣
 ↪else 0

            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)
```



[ ]:
```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)
```

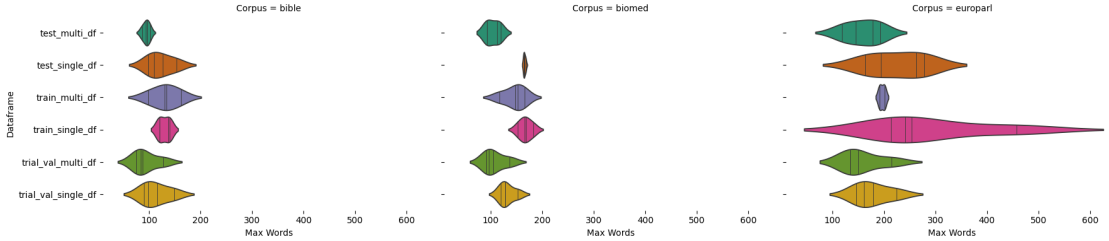
```
        def compute_span_metrics_no_contracts(sentence):
            if pd.isna(sentence):
                return pd.Series({'word_count': 0, 'char_count': 0,␣
↪'avg_word_len': 0})

            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_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)
```



[ ]:
```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:
```

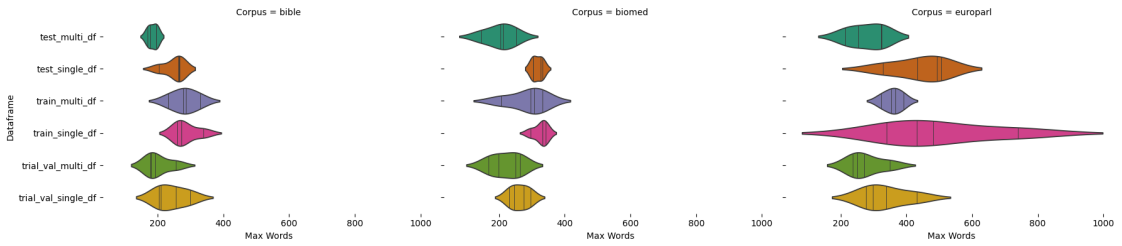
```
                return 'Q4'

        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_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)
```



[ ]:
```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:
```

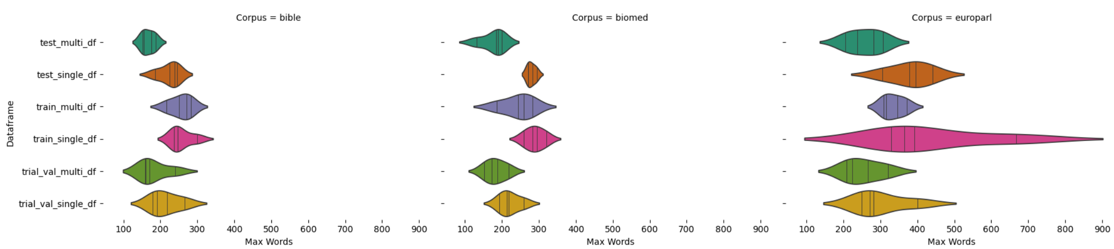
```
            return 'Q2'
        elif x <= q3:
            return 'Q3'
        else:
            return 'Q4'

    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_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)
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)
```

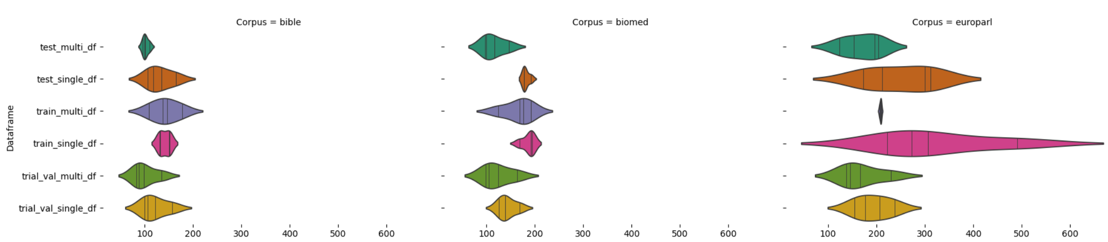
```
        def get_quartile(x):
            if x <= q1:
                return 'Q1'
            elif x <= q2:
                return 'Q2'
            elif x <= q3:
                return 'Q3'
            else:
                return 'Q4'

        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}")

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()
```

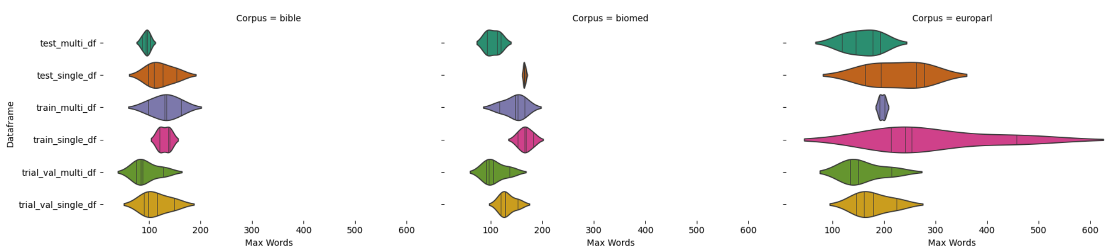
```
        q1 = df['complexity'].quantile(0.25)
        q2 = df['complexity'].quantile(0.50)
        q3 = df['complexity'].quantile(0.75)

        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)
```



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

def analyze_sentence_spans_by_corpus_and_quartile_no_contracts(dfs_dict):
    results = []
```

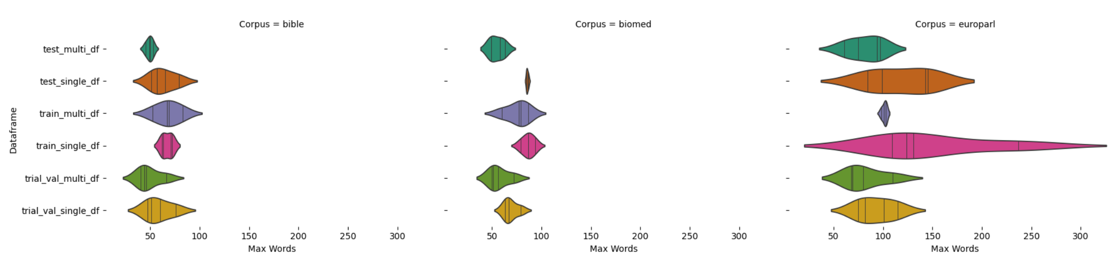
```
    for df_name, df in dfs_dict.items():
        print(f"Processing {df_name} on 'newly created columns'...")
        df = df.copy()

        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_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
```

[ ]:
```
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"
]

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: (7000, 21)
binary_complexity
0    3534
1    3466
Name: count, dtype: int64

>>> train_multi_df shape: (1300, 21)
binary_complexity
0    665
1    635
Name: count, dtype: int64

>>> trial_val_single_df shape: (1000, 21)
binary_complexity
0    518
1    482
Name: count, dtype: int64

>>> trial_val_multi_df shape: (250, 21)
binary_complexity
0    142
1    108
Name: count, dtype: int64

>>> test_single_df shape: (1000, 21)
binary_complexity
0    518
1    482
Name: count, dtype: int64

>>> test_multi_df shape: (250, 21)
```

```
binary_complexity
1    127
0    123
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
```

[ ]:
```

```