Initial Results from DH Analysis on Newspaper Novels and txtLAB corpora[[1]](#footnote-1)

Contents

[Stats Analysis over Time 2](#_Toc98241357)

[Word Counts over Time 2](#_Toc98241358)

[Sentences Counts over Time 3](#_Toc98241359)

[Average Words per Sentence over Time 4](#_Toc98241360)

[TTR Analysis over Time 5](#_Toc98241361)

[TTR Analysis over Time (Post 1814 novels) 7](#_Toc98241362)

[Percentage of Parts of Speech Counts in NNOVELS over Time 9](#_Toc98241363)

[Percentage of Parts of Speech Counts in TXTLAB over Time 10](#_Toc98241364)

[%POS\_NOUNS in NNOVELS and TXTLAB over Time 11](#_Toc98241365)

[%POS\_VERBS in NNOVELS and TXTLAB over Time 12](#_Toc98241366)

[%POS\_ADJECTIVES in NNOVELS and TXTLAB over Time 13](#_Toc98241367)

[%POS\_ADVERBS in NNOVELS and TXTLAB over Time 14](#_Toc98241368)

[Scatter Plots of Stats and POS 15](#_Toc98241369)

[Plotting %POS\_NOUNS to %POS\_VERBS for NNOVELS and TXTLAB 15](#_Toc98241370)

[Plotting %POS\_ADJS to %POS\_ADVS for NNOVELS and TXTLAB 16](#_Toc98241371)

[Plotting %POS\_PRON to %POS\_PROPN for NNOVELS and TXTLAB 17](#_Toc98241372)

[Pie Charts 18](#_Toc98241373)

[NNOVELS 18](#_Toc98241374)

[TXTLAB 19](#_Toc98241375)

# Stats Analysis over Time

## Chart, line chart, histogram Description automatically generatedWord Counts over Time

['The AVG of words\_count\_stopless of NNOVELS is 163686.7']

['The MIN of words\_count\_stopless of NNOVELS is 24909.0']

['The MAX of words\_count\_stopless of NNOVELS is 354478.0']

['The AVG of words\_count\_stopless of TXTLAB is 123274.19333333333']

['The MIN of words\_count\_stopless of TXTLAB is 23275.0']

['The MAX of words\_count\_stopless of TXTLAB is 356432.0']

**Notes:**

Note that the word count in the NNOVELS corpus is significantly higher during the period in which Agony Columns were popular, c.1840-1880.

## Sentences Counts over Time

Graphical user interface, chart, line chart

Description automatically generated

['The AVG of sentences\_count of NNOVELS is 10631.971428571429']

['The MIN of sentences\_count of NNOVELS is 1909.0']

['The MAX of sentences\_count of NNOVELS is 24438.0']

['The AVG of sentences\_count of TXTLAB is 7264.133333333333']

['The MIN of sentences\_count of TXTLAB is 1058']

['The MAX of sentences\_count of TXTLAB is 26855']

Notes:

Like the word counts, the sentences counts in the NNOVELS corpus are significantly higher during the period in which Agony Columns were popular, c.1840-1880.

## Chart, line chart Description automatically generatedAverage Words per Sentence over Time

['The AVG of average\_words\_per\_sentence of NNOVELS is 15.844283556728685']

['The MIN of average\_words\_per\_sentence of NNOVELS is 11.19336544911275']

['The MAX of average\_words\_per\_sentence of NNOVELS is 24.41286589516678']

['The AVG of average\_words\_per\_sentence of TXTLAB is 17.87748068561478']

['The MIN of average\_words\_per\_sentence of TXTLAB is 9.317274684354452']

['The MAX of average\_words\_per\_sentence of TXTLAB is 42.52860858257477']

**Notes:**

Despite the difference in word and sentences counts (meaning the books were longer in NNOVELS), the average words per sentence in the NNOVELS corpus is lower than that of the TXTLAB corpus. This may suggest that, during the popular period of Agony Columns, the NNOVELs became more concise with their sentences, perhaps resembling the short, declarative sentences of the newspapers at the time.

## TTR Analysis over Time

['The AVG of full\_text\_ttr of NNOVELS is 0.08290950030189044']

['The MIN of full\_text\_ttr of NNOVELS is 0.0392220390505141']

['The MAX of full\_text\_ttr of NNOVELS is 0.1610358294236364']

['The AVG of full\_text\_ttr of TXTLAB is 0.09223837352968185']

['The MIN of full\_text\_ttr of TXTLAB is 0.0366031633337277']

['The MAX of full\_text\_ttr of TXTLAB is 0.1993181318278466']

['The AVG of mattr\_500 of NNOVELS is 0.5150712853537941']

['The MIN of mattr\_500 of NNOVELS is 0.4597617454423514']

['The MAX of mattr\_500 of NNOVELS is 0.5637052983994336']

['The AVG of mattr\_500 of TXTLAB is 0.5179328530500183']

['The MIN of mattr\_500 of TXTLAB is 0.3673956472953637']

['The MAX of mattr\_500 of TXTLAB is 0.5771000303282796']

['The AVG of mattr\_2000 of NNOVELS is 0.3628188536623376']

['The MIN of mattr\_2000 of NNOVELS is 0.3019184792873604']

['The MAX of mattr\_2000 of NNOVELS is 0.4306457705131664']

['The AVG of mattr\_2000 of TXTLAB is 0.365972070932787']

['The MIN of mattr\_2000 of TXTLAB is 0.2133159679261588']

['The MAX of mattr\_2000 of TXTLAB is 0.4451271294036951']

**Notes:**

The TXTLAB corpus scored higher for all TTR analysis outputs (meaning, full text TTR analysis, as well as for computing the average TTR over smaller windows of 500 and 2000 words respectively), suggesting that the level of lexical diversity (or text complexity) is higher in the general 19th century corpus. However, the graph reveals that much of this “complexity” comes from novels that were published before the beginning of our corpus (in 1814). The subsequent visualization shows the same graph but with only with novels from 1814 onwards.

## Chart Description automatically generatedTTR Analysis over Time (Post 1814 novels)

['The AVG of full\_text\_ttr of NNOVELS is 0.08290950030189044']

['The MIN of full\_text\_ttr of NNOVELS is 0.0392220390505141']

['The MAX of full\_text\_ttr of NNOVELS is 0.1610358294236364']

['The AVG of full\_text\_ttr of TXTLAB is 0.09253788895627311']

['The MIN of full\_text\_ttr of TXTLAB is 0.0385683574458195']

['The MAX of full\_text\_ttr of TXTLAB is 0.1993181318278466']

['The AVG of mattr\_500 of NNOVELS is 0.5150712853537941']

['The MIN of mattr\_500 of NNOVELS is 0.4597617454423514']

['The MAX of mattr\_500 of NNOVELS is 0.5637052983994336']

['The AVG of mattr\_500 of TXTLAB is 0.5172929984214284']

['The MIN of mattr\_500 of TXTLAB is 0.3673956472953637']

['The MAX of mattr\_500 of TXTLAB is 0.5771000303282796']

['The AVG of mattr\_2000 of NNOVELS is 0.3628188536623376']

['The MIN of mattr\_2000 of NNOVELS is 0.3019184792873604']

['The MAX of mattr\_2000 of NNOVELS is 0.4306457705131664']

['The AVG of mattr\_2000 of TXTLAB is 0.3650274401096196']

['The MIN of mattr\_2000 of TXTLAB is 0.2133159679261588']

['The MAX of mattr\_2000 of TXTLAB is 0.4451271294036951']

**Notes:**

The TXTLAB corpus still scores higher for all TTR scores, although the comparison is very close. It appears that the level of lexical diversity is very similar across the novels in both corpora in this time frame. However, there are two interesting things to note here. 1) The NNOVELS corpus has more consistency, meaning the graph lines do not fluctuate as radically as the TXTLAB corpus. 2) Although the TXTLAB corpus scored higher in the average, the NNOVELS corpus consistently has a higher MIN and a lower MAX value than the TXTLAB corpus (this is also true for all the above analyses, such as word and sentences counts.) What this may suggest is that the NNOVEL corpus is more similar to itself, a criterion that would make sense for a specific literary genre.

## Percentage of Parts of Speech Counts in NNOVELS over Time

Chart, line chart

Description automatically generated

['The AVG of %POS\_NOUN of NNOVELS is 15.637814285714292']

['The MIN of %POS\_NOUN of NNOVELS is 12.565']

['The MAX of %POS\_NOUN of NNOVELS is 20.808']

['The AVG of %POS\_VERB of NNOVELS is 13.724514285714285']

['The MIN of %POS\_VERB of NNOVELS is 10.649']

['The MAX of %POS\_VERB of NNOVELS is 15.385']

['The AVG of %POS\_ADJ of NNOVELS is 6.2335714285714285']

['The MIN of %POS\_ADJ of NNOVELS is 4.709']

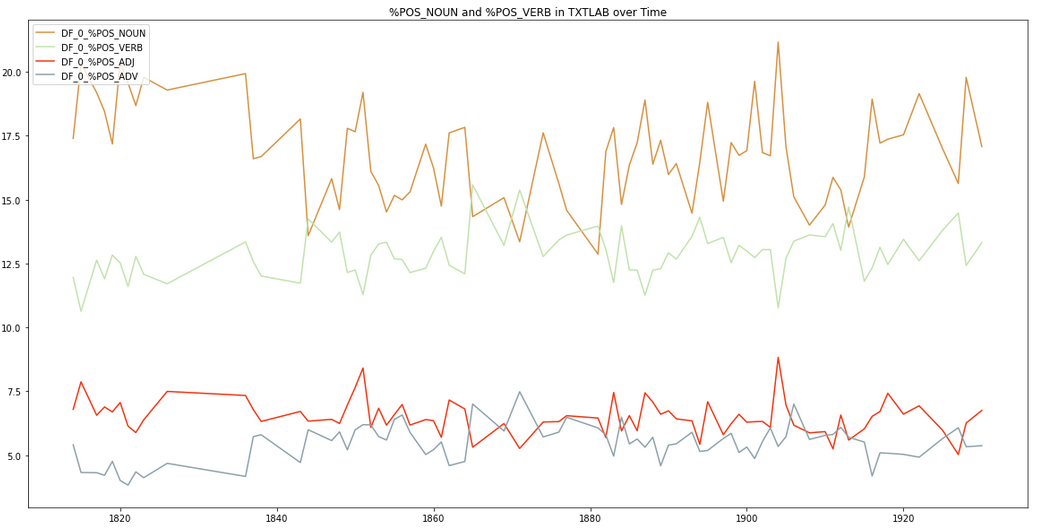
['The MAX of %POS\_ADJ of NNOVELS is 8.54']

['The AVG of %POS\_ADV of NNOVELS is 5.8599000000000006']

['The MIN of %POS\_ADV of NNOVELS is 4.536']

['The MAX of %POS\_ADV of NNOVELS is 7.528']

## Percentage of Parts of Speech Counts in TXTLAB over Time



['The AVG of %POS\_NOUN of TXTLAB is 16.721362903225803']

['The MIN of %POS\_NOUN of TXTLAB is 10.263']

['The MAX of %POS\_NOUN of TXTLAB is 21.156']

['The AVG of %POS\_VERB of TXTLAB is 12.911717741935488']

['The MIN of %POS\_VERB of TXTLAB is 10.63']

['The MAX of %POS\_VERB of TXTLAB is 15.574']

['The AVG of %POS\_ADJ of TXTLAB is 6.527088709677421']

['The MIN of %POS\_ADJ of TXTLAB is 4.7']

['The MAX of %POS\_ADJ of TXTLAB is 8.847']

['The AVG of %POS\_ADV of TXTLAB is 5.492064516129033']

['The MIN of %POS\_ADV of TXTLAB is 3.763']

['The MAX of %POS\_ADV of TXTLAB is 10.747']

**Notes:**

The two corpora are quite similar in terms of POS tagging, at least at the level of nouns, verbs, adjectives, and adverbs. However, there are slight differences. 1) The NNOVELS corpus scored slightly higher than the TXTLAB corpus in terms of verb% (NNOVELS = 13.725%; TXTLAB = 12.912%), and scored slightly lower in terms of noun% (NNOVELS = 15.638%; TXTLAB = 16.721%). 2) Following the same pattern, the NNOVELS corpus scored slightly higher than the TXTLAB corpus in terms of adverbs% (NNOVELS = 5.860%; TXTLAB = 5.492%), and scored slightly lower in terms of adjectives% (NNOVELS = 6.234%; TXTLAB = 6.530%). These statistics may suggest a slight difference in “action” aesthetics, and that there occurs more movement or activity in the NNOVELS corpus (as a result of more verbs and adverbs compared to nouns and adjectives). 3) As before, the NNOVELS corpus is much more similar to itself, exhibiting higher MIN and lower MAX scores across the board.

The next 4 graphs represent the same information, but compare POS tags of both corpora and filtering the max years to match (1814 < x > 1916).

## %POS\_NOUNS in NNOVELS and TXTLAB over Time

Chart, line chart, histogram

Description automatically generated

['The AVG of %POS\_NOUN of NNOVELS is 15.637814285714292']

['The MIN of %POS\_NOUN of NNOVELS is 12.565']

['The MAX of %POS\_NOUN of NNOVELS is 20.808']

['The AVG of %POS\_NOUN of TXTLAB is 16.613954954954952']

['The MIN of %POS\_NOUN of TXTLAB is 10.263']

['The MAX of %POS\_NOUN of TXTLAB is 21.156']

## %POS\_VERBS in NNOVELS and TXTLAB over Time

['The AVG of %POS\_VERB of NNOVELS is 13.724514285714285']

['The MIN of %POS\_VERB of NNOVELS is 10.649']

['The MAX of %POS\_VERB of NNOVELS is 15.385']

['The AVG of %POS\_VERB of TXTLAB is 12.87960360360361']

['The MIN of %POS\_VERB of TXTLAB is 10.63']

['The MAX of %POS\_VERB of TXTLAB is 15.574']

## Chart, line chart, histogram Description automatically generated%POS\_ADJECTIVES in NNOVELS and TXTLAB over Time

['The AVG of %POS\_ADJ of NNOVELS is 6.2335714285714285']

['The MIN of %POS\_ADJ of NNOVELS is 4.709']

['The MAX of %POS\_ADJ of NNOVELS is 8.54']

['The AVG of %POS\_ADJ of TXTLAB is 6.523333333333336']

['The MIN of %POS\_ADJ of TXTLAB is 4.7']

['The MAX of %POS\_ADJ of TXTLAB is 8.847']

## %POS\_ADVERBS in NNOVELS and TXTLAB over Time

Chart, histogram

Description automatically generated

['The AVG of %POS\_ADV of NNOVELS is 5.8599000000000006']

['The MIN of %POS\_ADV of NNOVELS is 4.536']

['The MAX of %POS\_ADV of NNOVELS is 7.528']

['The AVG of %POS\_ADV of TXTLAB is 5.524792792792794']

['The MIN of %POS\_ADV of TXTLAB is 3.763']

['The MAX of %POS\_ADV of TXTLAB is 10.747']

**Notes:**

Several interesting things to note while looking at these graphs: 1) The NNOVELS are much more consistent between 1860-1880, possibly suggesting stronger similarities within the corpus in this range. 2) The few novels we have from before 1840 and those after 1900 in both corpora tend to be drastically different from each other. In a way, it seems as if they started out very different and become more similar in the period of the 1840s-1900, after which they depart in two different directions. 3) Perhaps most interesting is the relative mirroring that happens between the two corpora. This is especially evident in the NOUNS and VERBS charts, where each corpus seems to follow the exact opposite trend of the other; in the years in which the TXTLAB corpus has a lower percentage of nouns or verbs, the NNOVELS spikes in those categories, and vice versa. It almost seems as they represent opposite stylistic techniques, published at the same moment in time (or perhaps, different reactions to the same things that were happening).

# Scatter Plots of Stats and POS

## Plotting %POS\_NOUNS to %POS\_VERBS for NNOVELS and TXTLAB

**Notes:**

This plot is not over time, but rather the Nouns/Verb axis for both corpora. Note how NNOVELS is much more uniform and packed together and follows the trend. As the NOUN% is higher, the VERB% is lower. This is not necessarily true for the TXTLAB corpus, which is a bit all over the place. The names appear for those in the radical 2% percentile, meaning in the lower and top 2% of either nouns, verbs or both. The chart below does the same thing for ADJ% and ADV%.

## Chart, scatter chart Description automatically generatedPlotting %POS\_ADJS to %POS\_ADVS for NNOVELS and TXTLAB

**Notes:**

As before, the NNOVELS corpus is much more uniform, with less outliers. There isn’t much of a trend happening here except for the observation that more of the ADV/ADJ doesn’t necessarily mean less of the other.

## Plotting %POS\_PRON to %POS\_PROPN for NNOVELS and TXTLAB

Chart, scatter chart

Description automatically generated

**Notes:**

As before, the NNOVELS corpus is much more uniform, with less outliers. But here there is definitely a trend, at least within the NNOVELS corpus. As pronouns go up, the proper nouns go down. However, the NNOVELS corpus contains much less pronouns than the TXTLAB corpus. This may be a possibly indicator of similarity to newspaper stylistics, which tend to use on the proper nouns of the object of their reportage. We could also check this using the same strategy on newspapers. It is also important to note that the same novels return as the outliers: *The Mirror of the Sea, Cecil, East Lynne, The Papers.*

# Pie Charts

## NNOVELS

Chart, pie chart

Description automatically generated

## TXTLAB

Chart, pie chart

Description automatically generated

**Notes:**

This is basically a static representation of the same data, but it reveals a few interesting things.

1) The NNOVEL corpus has less AUXiliaries overall, probably as a result of higher VERB counts.

2) The NNOVEL corpus simultaneously has a higher % of PROPNouns and a lower % of PRONouns, suggesting that the characters and locales are more defined within the NNOVELS corpus compared to the TXTLAB corpus

1. ## Remember that in all visualizations: DH\_0 = NNOVELS; DH\_1 = TXTLAB [↑](#footnote-ref-1)