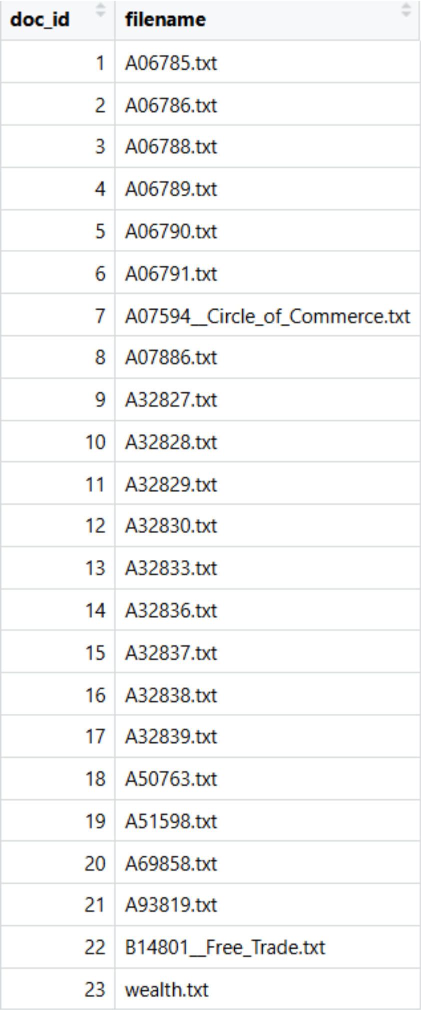
Data Exploration

My homework is based on the structured one.

For easier reference, I have renamed these documents to Text 1, Text 2, …, Text 23, and the corresponding table is shown below:



doc_id	filename
1	A06785.txt
2	A06786.txt
3	A06788.txt
4	A06789.txt
5	A06790.txt
6	A06791.txt
7	A07594__Circle_of_Commerce.txt
8	A07886.txt
9	A32827.txt
10	A32828.txt
11	A32829.txt
12	A32830.txt
13	A32833.txt
14	A32836.txt
15	A32837.txt
16	A32838.txt
17	A32839.txt
18	A50763.txt
19	A51598.txt
20	A69858.txt
21	A93819.txt
22	B14801__Free_Trade.txt
23	wealth.txt

Data Clean

During my data cleaning process at first, I only normalized all the long S characters, and then only regularly removed all the punctuations, numbers, symbols, and the stop words (the stop words list comes from our tutorial section, which includes the traditional English words such as "vnto", "hee", "beene", etc.). I decide not to normalize other spelling and typographical issues at first, because I'm personally unfamiliar with these raw texts, I want to learn more about them cautiously. I decided to preserve the original text as much as possible to prevent my human intervention.

After I computed my TF-IDF, I immediately read the top 15 words:

Firstly, I found there were some variant spellings "2dly" and "vpon", thus I returned to normalize "2dly" to "secondly", "vpon" to "upon";

Secondly, I noticed there are some isolated letters, such as "l", "p", or "s", significantly impacting the results. I returned to conduct a removal of all the isolated letters (some important letters with meanings, such as "I" or "a", have already been automatically removed by the stop words list “en”).

Thirdly, I also noticed the uncertain abbreviations “ll” and “ss” in Text 2 and Text 8. Because I’m unfamiliar with the abbreviations “ll” and “ss”, I use the function “kwic (toks, pattern = "ll")” to check their contexts and found it is probably an incorrect OCR output (for example, I found a sentence like “ll ll collen ll ausborgh ll munchen ll wisell ll norlingen” from Text 2). I decided to remove them.

Approach 1 TF–IDF: lexical distinctiveness

TF–IDF helps identify what makes each document distinctive relative to the corpus as a whole. I constructed the document–feature matrix (DFM) for the full corpus and recomputed the TF-IDF weights after cleaning the texts. I also extract the top 15 TF-IDF terms from each document (see Appendix 1).

◆ Do some documents share distinctive vocabulary?

Yes, there are some documents that share distinctive vocabulary. For example, Texts 1, 2, 3, 4, 6, and 19 share the term “moneys”; or Texts 9, 10, 13, and 15 share the term “cent”; or Texts 9, 13, and 15 share the term “interest”.

Because TF-IDF’s computation is based on the cross-text comparison, I think these shared distinctive terms could be a signal which reveal a potentially higher similarity between these specific texts’ content, rather than the commonalities among all the texts.

◆ Are distinctive terms topical, rhetorical, or technical?

These distinctive terms are different.

Some distinctive terms are topical, such as: moneys, interest, muslins, india, bombay, cloth, cent, silk, election, war, tax, etc.

Some distinctive terms are rhetorical: purely, richest, nearly, weight, unto, vs, etc.

Some are technical: Dorchester, factor, exchange, realm, etc.

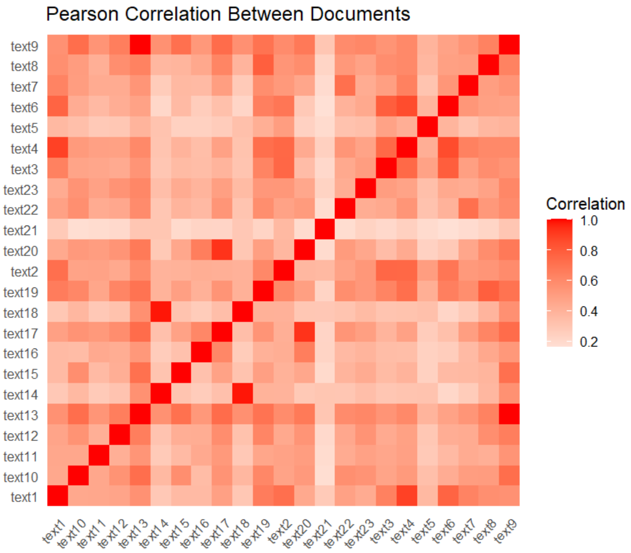
◆ Are there documents whose “distinctiveness” seems driven by noise or formatting rather than content?

To address this question, I focused on the top 3 words (which has the highest TF-IDF scores and impact each documents) of each documents to check each document again. I find for example, Text 2 (“hundreth”, “ounces”), Text 5 (“maketh”), Text 9 (“abatement”), Text 12 (“em”), and Text 15 (“user”, “users”), seem are driven by noise or formatting rather than content. Although these terms also suggest they are very distinctive from each document, they are hard for scholars to interpret and conduct a further content or topic analysis.

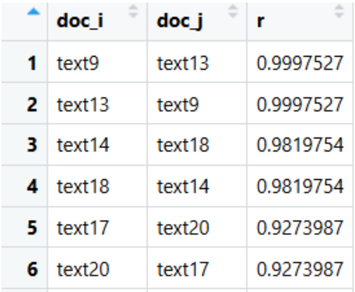
Approach 2 — Pearson correlation: similarity and distance between text

Pearson correlation treats each document as a vector and asks: which texts behave similarly across the vocabulary space?

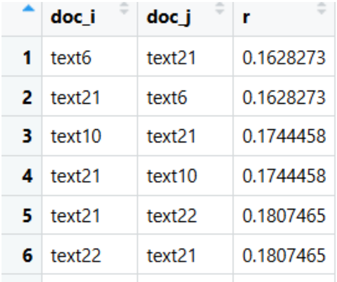
I first trim the very rare words from the DFM to decline its potential distraction. Then, I used the DFM to compute pairwise Pearson correlations between document. I first convert the correlation matrix into a long format data frame, and then find the most and least similar pairs. Finally, I visualize the results on a similarity heatmap:



- Most and Least Similar Paris:
  The most similar document pairs: Text 9 and 13; Text 14 and18.
  The least similar document pairs: Text 6 and 21; Text 10 and 21.



	doc_i	doc_j	r
1	text9	text13	0.9997527
2	text13	text9	0.9997527
3	text14	text18	0.9819754
4	text18	text14	0.9819754
5	text17	text20	0.9273987
6	text20	text17	0.9273987



	doc_i	doc_j	r
1	text6	text21	0.1628273
2	text21	text6	0.1628273
3	text10	text21	0.1744458
4	text21	text10	0.1744458
5	text21	text22	0.1807465
6	text22	text21	0.1807465

- What questions would you ask from this corpus after seeing the patterns of similarity/differences that you see here? Explain your answer

1. My first question is why Text 9 and Text 13, and also Text 14 and Text 18 are so similar to each other.

   From the method perspective, during the TF-IDF analysis before, I have found Texts 9 and 13 shared the same 10 terms in each of their top 15 lists, while Text 14 and 18 also have many same words. The Pearson correlation analysis has further suggested their similarity at the vocabulary level. These two analysis complement each other and both suggest the similarity between the two texts mathematically.

   However, when I notice their similarities have already closed to 1, it could be a warning and it is quite urgent and necessary for scholars to recheck their texts manually. Such a high similarity might suggest they are mistakenly analyzing the "same" text. For example, if scholars are processing the contents on social media, there might be a situation where some users copied or forwarded a post originally from another user but scholars who scraped them by machine have not yet reviewed them in person. They may need to remove some highly similar texts and even redo the analysis.

2. My second question is why Text 6 and Text 21, and also Text 10 and Text 21 are so different from each other.

   When dealing with the most distant/different text pairs in the same corpus, I think they are more important and informative for the textual analysis than the most similar text pairs.

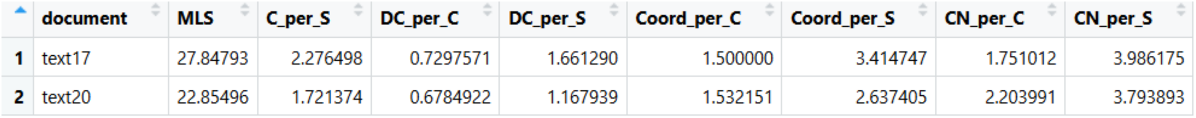
Differences may suggest either the most disparate linguistic styles (different authors, eras, regions, or genres may all impact their languages), or the most distinct text topics and contents. What factors cause their differences? I think further analyze them can significantly enrich scholars' understanding of this corpus, as well as propose the research questions. Meanwhile, if some texts are too different from the others, it could also be a signal for researchers to consider if they may choose some inappropriate texts for their topic.

Approach 3 — Syntactic complexity profile

Syntactic complexity captures how texts are written, rather than what words they use.

Because the results in the TF-IDF weights and the Pearsons Correlation, and also my personally reading experience, I have noticed Text 9 and Text 13, and also Text 14 and Text 18 are almost the same text since they are too similar to be analyzable. I decide to use the third highest similar text pairs, Text 17 and Text 20 (r=0.927), to analyze the syntactic complexity. I want to further understand how these two texts similar or different from each other.

In this section, I first measured sentence length, clause counts, dependent clauses, coordination and complex nominals of the Text 17 and Text 20. Then, I computed the mean length of sentence (MLS), clauses per sentence (C_per_S), dependent clauses per clause and/or sentence (DC_per_C/DC_per_S), coordination per clause and/or sentence (Coord_per_C/ Coord_per_S), complex nominals per clause and/or sentence (CN_per_C/CN_per_S) with the result below table and visualization:



	document	MLS	C_per_S	DC_per_C	DC_per_S	Coord_per_C	Coord_per_S	CN_per_C	CN_per_S
1	text17	27.84793	2.276498	0.7297571	1.661290	1.500000	3.414747	1.751012	3.986175
2	text20	22.85496	1.721374	0.6784922	1.167939	1.532151	2.637405	2.203991	3.793893

Table 3 1

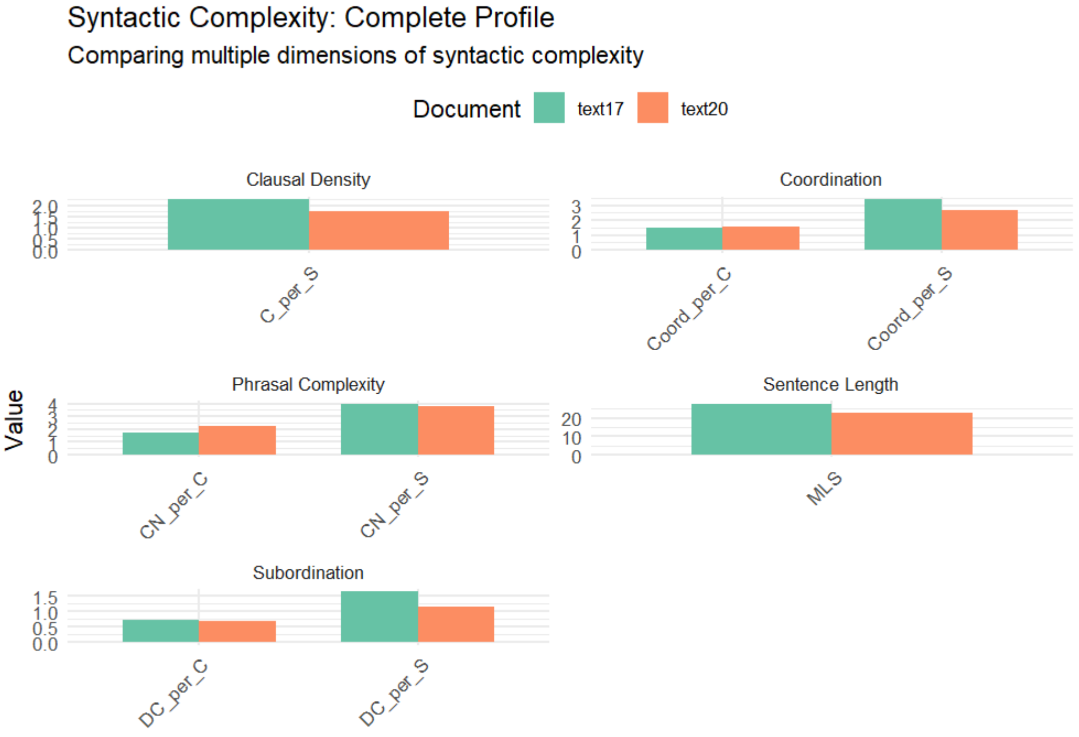


Table 3 2

- ◆ How do the two texts differ in syntactic complexity?

In terms of the results above, I found there is also a high Syntactic similarity between Text 17 and Text 20, but the analysis of their syntactic complexity also suggest their nuanced differences in language styles. Text 17 has a longer mean length of sentence and more sentences and clauses, which may cause it have more coordination, complex nominals, and dependent clauses per sentence, and suggest a more complex grammar structure than Text 20. Text 20, with a shorter mean length of sentence and less clauses, probably exhibits an easier language styles for audiences to read.

I randomly picked two real sentence examples from Text 17 and Text 20, respectively, to illustrate this difference straightforwardly (see below). The example sentence from Text 17 is longer and has a very complex structure, whereas the example from Text 20 uses fewer clauses with simpler dependency relation

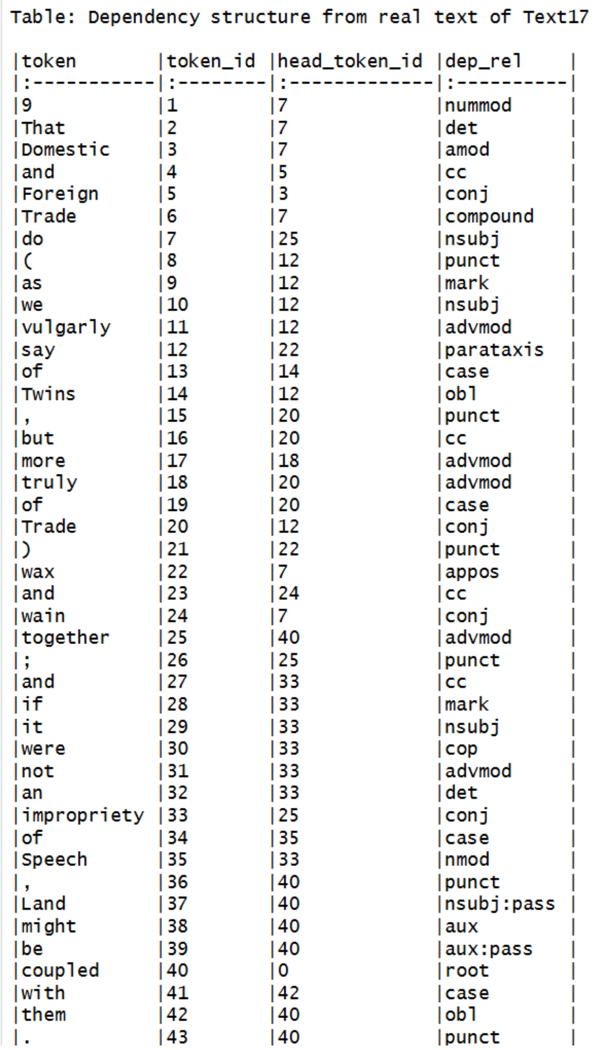


Table: Dependency structure from real text of Text17

token	token_id	head_token_id	dep_rel
9	1	7	nummod
That	2	7	det
Domestic	3	7	amod
and	4	5	cc
Foreign	5	3	conj
Trade	6	7	compound
do	7	25	nsubj
(	8	12	punct
as	9	12	mark
we	10	12	nsubj
vulgarly	11	12	advmod
say	12	22	parataxis
of	13	14	case
Twins	14	12	obl
,	15	20	punct
but	16	20	cc
more	17	18	advmod
truly	18	20	advmod
of	19	20	case
Trade	20	12	conj
)	21	22	punct
wax	22	7	appos
and	23	24	cc
wain	24	7	conj
together	25	40	advmod
;	26	25	punct
and	27	33	cc
if	28	33	mark
it	29	33	nsubj
were	30	33	cop
not	31	33	advmod
an	32	33	det
impropriety	33	25	conj
of	34	35	case
Speech	35	33	nmod
,	36	40	punct
Land	37	40	nsubj:pass
might	38	40	aux
be	39	40	aux:pass
coupled	40	0	root
with	41	42	case
them	42	40	obl
.	43	40	punct

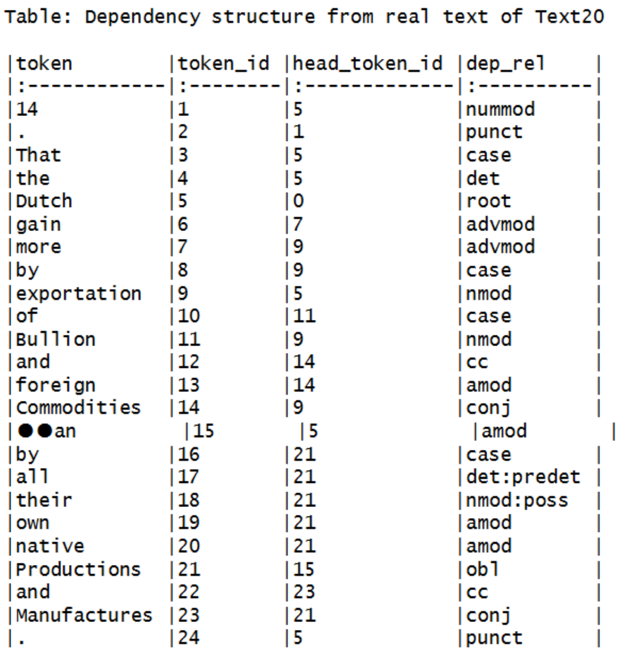


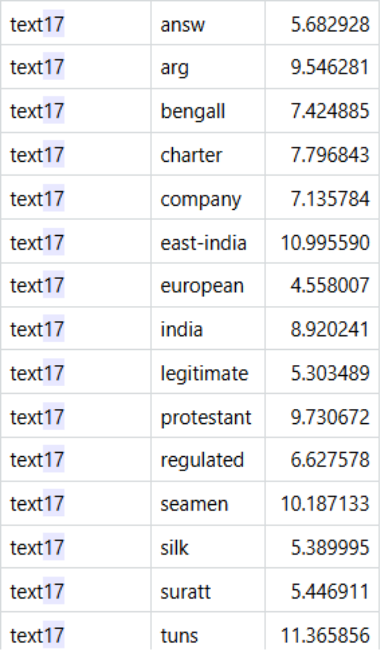
Table: Dependency structure from real text of Text20

token	token_id	head_token_id	dep_rel
14	1	5	nummod
.	2	1	punct
That	3	5	case
the	4	5	det
Dutch	5	0	root
gain	6	7	advmod
more	7	9	advmod
by	8	9	case
exportation	9	5	nmod
of	10	11	case
Bullion	11	9	nmod
and	12	14	cc
foreign	13	14	amod
Commodities	14	9	conj
●●an	15	5	amod
by	16	21	case
all	17	21	det:predet
their	18	21	nmod:poss
own	19	21	amod
native	20	21	amod
Productions	21	15	obl
and	22	23	cc
Manufactures	23	21	conj
.	24	5	punct

◆ Do these differences align with or complicate your earlier lexical findings?

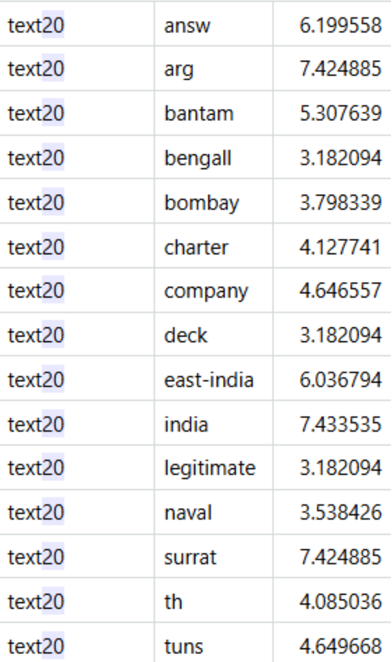
I think the results align with the earlier lexical findings and these two analyses could contemplate each other. Lexical complexity focuses on the diverse vocabulary that may suggest the content of texts, while the syntactic complexity further addresses the grammar structure reveal authors'/texts' language skills and styles.

This syntactic complexity has suggested Text 17 and Text 20 have a similar (with slight differences) language styles. When we recheck the Top 15 words in Text 17 and Text 20 of TF-IDF weights (as shown in the table below Table 3.3 and 3.4), we could find they also shared 10 top words that probably implies these two texts have similar (with slight difference) contents and topics. Both results indicate that Text 17 and Text 20 share a high degree of similarity, yet they are not identical at the same time.



text17	answ	5.682928
text17	arg	9.546281
text17	bengall	7.424885
text17	charter	7.796843
text17	company	7.135784
text17	east-india	10.995590
text17	european	4.558007
text17	india	8.920241
text17	legitimate	5.303489
text17	protestant	9.730672
text17	regulated	6.627578
text17	seamen	10.187133
text17	silk	5.389995
text17	suratt	5.446911
text17	tuns	11.365856

Table 3 3



text20	answ	6.199558
text20	arg	7.424885
text20	bantam	5.307639
text20	bengall	3.182094
text20	bombay	3.798339
text20	charter	4.127741
text20	company	4.646557
text20	deck	3.182094
text20	east-india	6.036794
text20	india	7.433535
text20	legitimate	3.182094
text20	naval	3.538426
text20	surrat	7.424885
text20	th	4.085036
text20	tuns	4.649668

Table 3 4

4. Synthesis — triangulating evidence

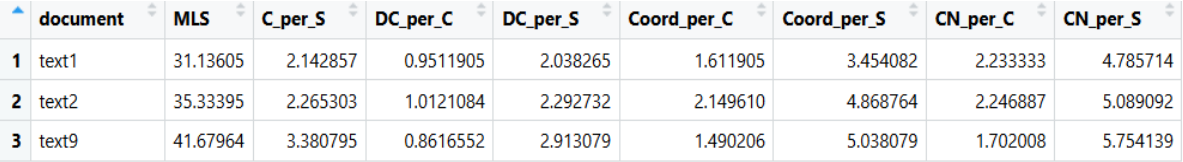
After computing the TF-IDF weights, Person Correlation, and Syntactic complexity of these 23 texts under the same topic of "political economics", I agree these results can help scholars to further explore relationships between these texts and also find the potential research questions.

◆ Central Analytical Question/Hypothesis/Idea

One hypothesis arises from these analyses is why some TF-IDF distinctive terms are rhetorical, or technical? Does that mean such texts possess a more complex syntactic style than others?

The TF-IDF weights expressed the relationship of Terms Frequence (TF) and Inverse Document Frequency (IDF), while a high score suggests one word appears frequently with a text and relatively rare/distinctive among all the texts. This may suggest how these rhetorical and technical TD-IDF words are frequently appears but also relatively distinctive.

To address this question, I specifically focused on the Top 3 TF-IDF distinctive terms from these 23 texts and then identify texts 1, 2, 9, 13, and 23 are all have distinctive rhetorical or technical terms. Since Text 23 is quite long (my laptop is unable to run syntactic analysis for it) and Text 9 and Text 13 are nearly same, I focus on the text 1,2,9 to analyze their syntactic complexity.



	document	MLS	C_per_S	DC_per_C	DC_per_S	Coord_per_C	Coord_per_S	CN_per_C	CN_per_S
1	text1	31.13605	2.142857	0.9511905	2.038265	1.611905	3.454082	2.233333	4.785714
2	text2	35.33395	2.265303	1.0121084	2.292732	2.149610	4.868764	2.246887	5.089092
3	text9	41.67964	3.380795	0.8616552	2.913079	1.490206	5.038079	1.702008	5.754139

Table 4 1

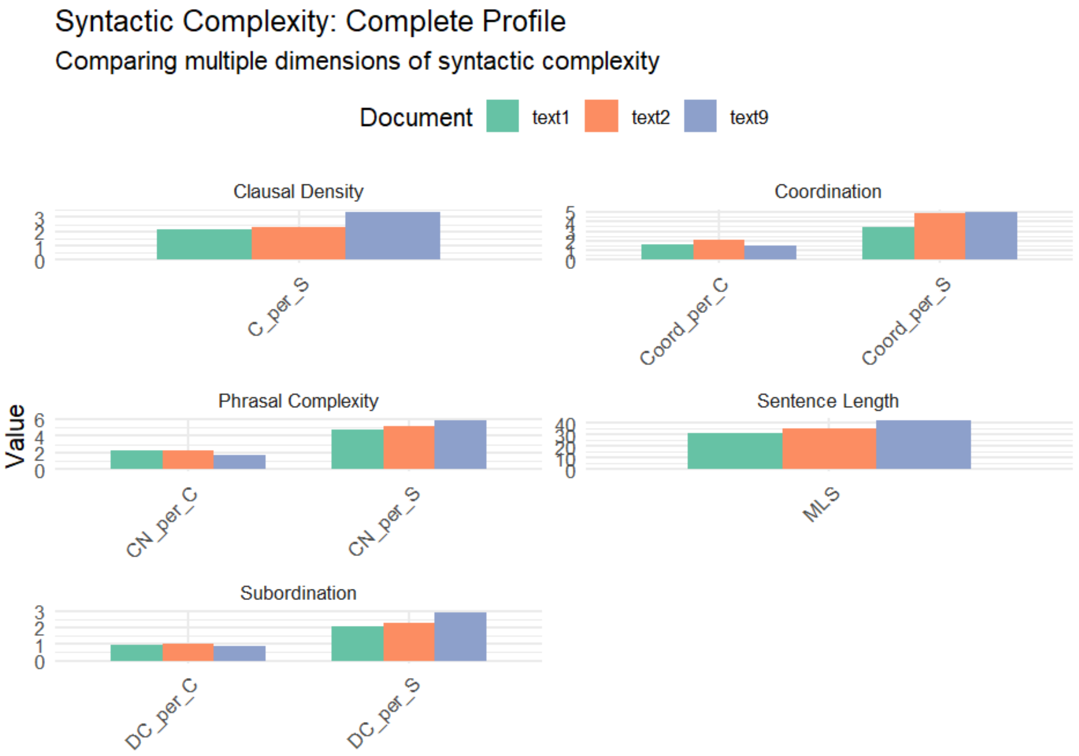


Table 4 2

In terms of the results in Tables 4.1 and 4.2, all three texts exhibit a longer mean length of sentences, dependent clauses per clause and/or sentence, coordination per clause and/or sentence, and complex nominals per clause and/or sentence, compared to the earlier results of Text 17 and Text 20. This may support the hypothesis that the texts with rhetorical, or technical TF-IDF terms probably possess a more complex syntactic style than others, but it still require further methodological examination to testify.

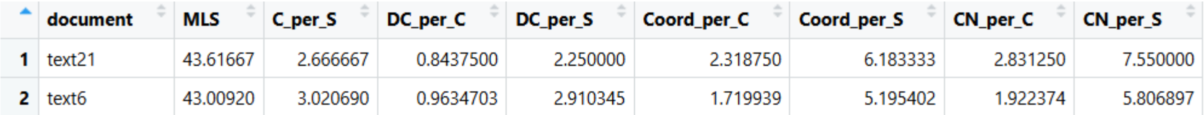
◆ Methodological Reflection: How do we determine the relationships between texts?

The relationship between these three analyses also raise another question: how can we use a methodological way to examine our corpus and remove inappropriate (repeated) text, especially when there is a huge number of texts that are automatically scraped without reviewing by human in personally.

By using Text 17 and Text 20 as examples, these two texts share 10 words among their top 15 TF-IDF weighted terms, which suggests that based on their lexical choices, their content and themes may be highly similar. During the Pearson Correlation analysis, both texts also suggest a high similarity level (r=0.927). In the syntactic analysis, both texts have also shown that they have highly similar language styles. These three analyses have all similarly suggested Text 17 and Text 20 would be highly similar to each other.

If Text 17 and Text 20 are one of the most similar pairs, then I reexamined the lexical and syntactic complexities of the least similar pairs Text 6 and Text 21 (see below Table 4.3 to 4.6). I found that the relationship between these three analyses began to become increasingly complicated: while Text 6 and Text 21 possess totally different distinctive words in the TF-IDF analysis suggested they may discuss very different contents/topics, they report a relatively similar syntactic expression.

In fact, these three results could contemplate each other but not be necessarily related to each other. How we define the relationship between these texts actually depend on the perspective and emphasis we choose, while scholars must weigh the different purposes of these methods separately. However, these three analyses may bring an "unexpected" function for scholars to check their corpus. If all three analyses consistently show a high degree of similarity, this could be a red flag that researchers should pay attention to, as it may indicate the corpus contain the repeated texts or the same texts' different versions.



	document	MLS	C_per_S	DC_per_C	DC_per_S	Coord_per_C	Coord_per_S	CN_per_C	CN_per_S
1	text21	43.61667	2.666667	0.8437500	2.250000	2.318750	6.183333	2.831250	7.550000
2	text6	43.00920	3.020690	0.9634703	2.910345	1.719939	5.195402	1.922374	5.806897

Table 4 3

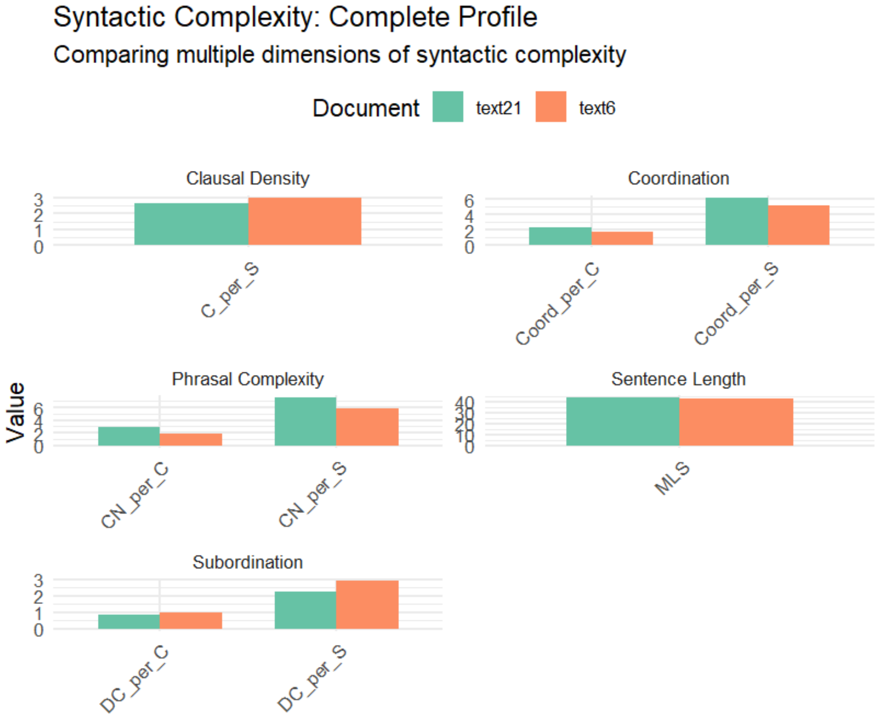
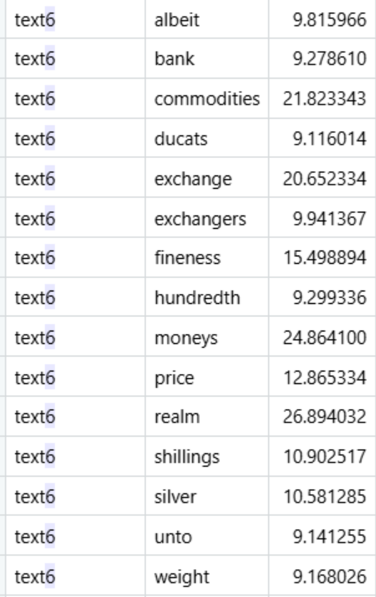
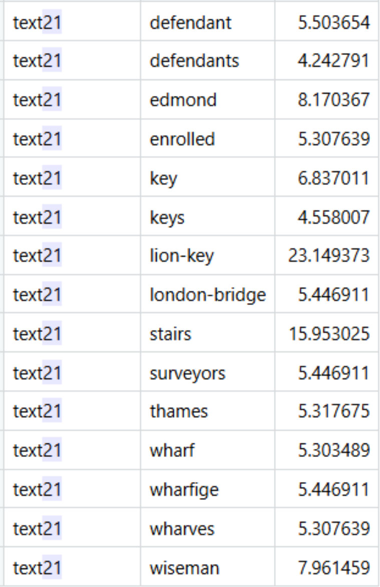


Table 4 4



text6	albeit	9.815966
text6	bank	9.278610
text6	commodities	21.823343
text6	ducats	9.116014
text6	exchange	20.652334
text6	exchangers	9.941367
text6	fineness	15.498894
text6	hundredth	9.299336
text6	moneys	24.864100
text6	price	12.865334
text6	realm	26.894032
text6	shillings	10.902517
text6	silver	10.581285
text6	unto	9.141255
text6	weight	9.168026

Table 4 5



text21	defendant	5.503654
text21	defendants	4.242791
text21	edmond	8.170367
text21	enrolled	5.307639
text21	key	6.837011
text21	keys	4.558007
text21	lion-key	23.149373
text21	london-bridge	5.446911
text21	stairs	15.953025
text21	surveyors	5.446911
text21	thames	5.317675
text21	wharf	5.303489
text21	wharfige	5.446911
text21	wharves	5.307639
text21	wiseman	7.961459

Table 4 6