Main, high level, research question:

How has both discourse and representation in the British parliament changed between 1803 and 2005 from the perspective of gender?

In this report, we do basic data analysis of the Hansard corpus. We first crawl the data (201 years worth, form 1803-2005, excepting 1816 and 1829 for which we did not find data), and parse it to create three different views of the metadata

1. Parquet files - contain debate and speaker data, separately for each year
2. Jsonline - contain all data parsed and arranged into jsonline per year, good for manual inspection
3. SQLite DB - untested so far - meant to enable keyword search

We first present some statistics around our data.

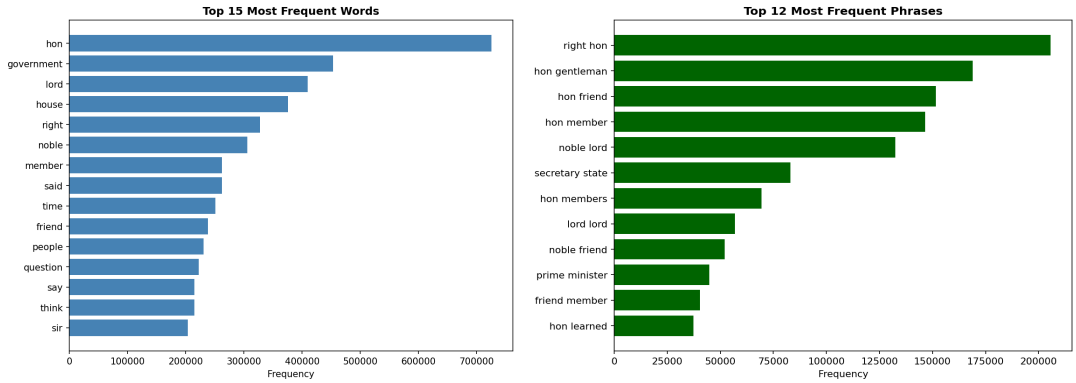
Core Statistics:
- Years: 201 years (1803-2005)
- Debates: 802,178 total debates
- Average Debate Length: 1,487 words
- Unique Speakers: 89,472 speakers
- Total Corpus: 1.19 billion words (~6.7 GB)
- Coverage: 99% of possible years (only missing 1816 & 1829)
- Chambers: 79.7% Commons, 20.3% Lords
- Speaker Activity: 29.9 average mentions per speaker
- Debates/Year: ~4,000 average

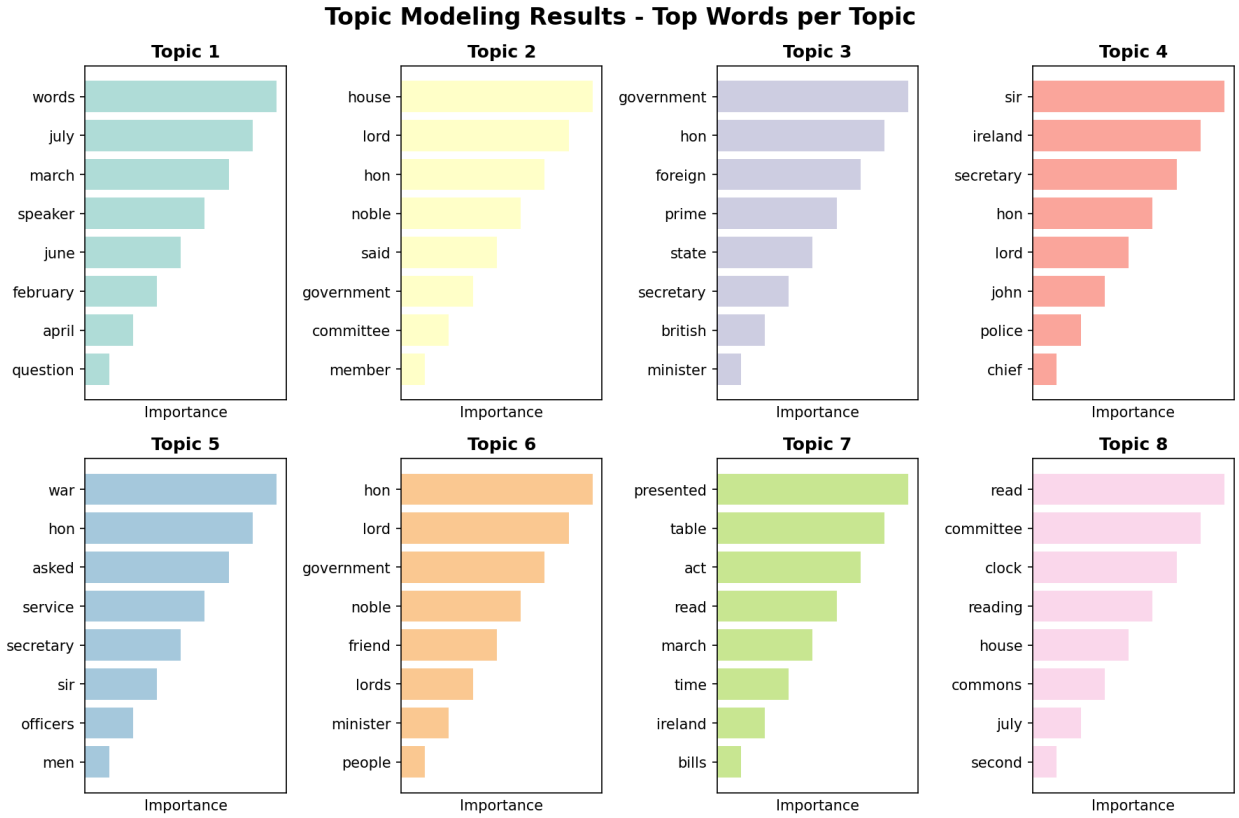
Overall, we see that this is a rich dataset with lots of valuable text data and associated metadata. Below, we attempt to analyze this data from the perspective of gender. We try to give a sense of overall distribution of the kind of data that we have collected, and then we show gender-based analysis.

Methodology:
For overall analysis, we sample 100k debates, using stratified sampling to give us a good representation for each year we have availability for, upon which we do simple unigram and bigram analysis, which show, as one would expect, a lot of procedural language which is very common in parliamentary speech. We also run topic modeling (using LDA) over the sample of 100k debates. In addition, we have speaker metadata, for which we use title-based heuristics (ex. Lord - Male, Duchess - Female) to identify gender.

Overall Corpus Analysis:

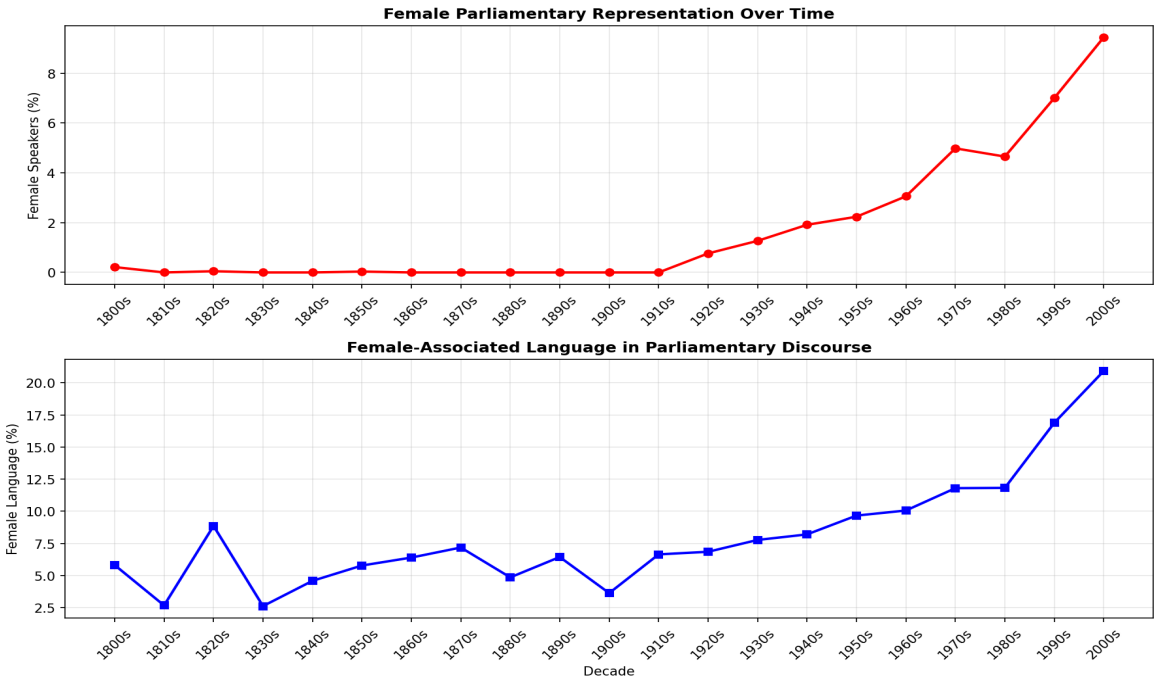
We see the words “lord”, “government”, “hon” - a contraction of the word honorable - are most common, which seems to follow the pattern that one would expect in parliament, since these are forms of addressing fellow members. Our bigram analysis also shows that formal forms of address are extremely common, again, due to the rule-based nature of parliament, where members address each other in a fairly standardized manner.

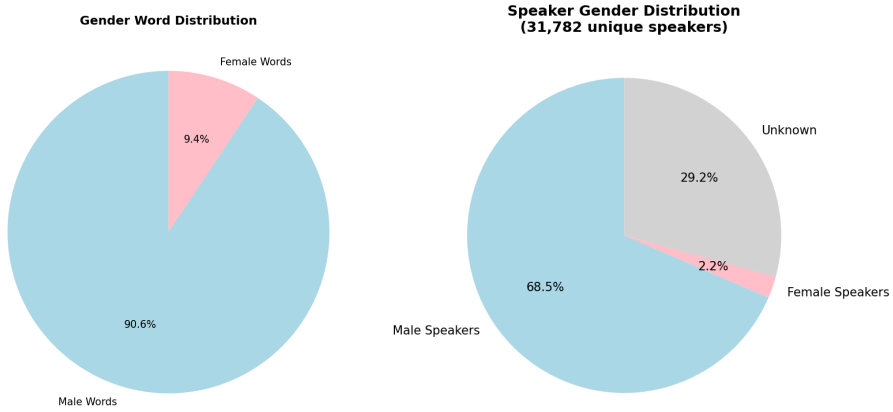




We see that there’s a steady increase in representation (starting in 1918), with an interesting drop in 1960 (check if this is noise or corresponds to other social events).

We also collect a set of gendered words (Zhao et. al. -2018) that we use to analyze the prevalence of gendered language over time. The female-associated language graph (the percentage is calculated over all gendered words in our word list) shows that there is an increased proportion of gendered words over time.

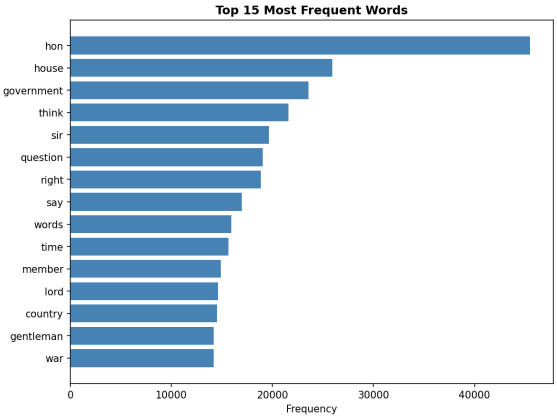


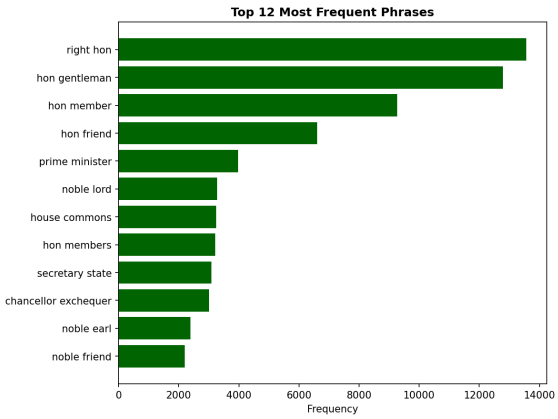


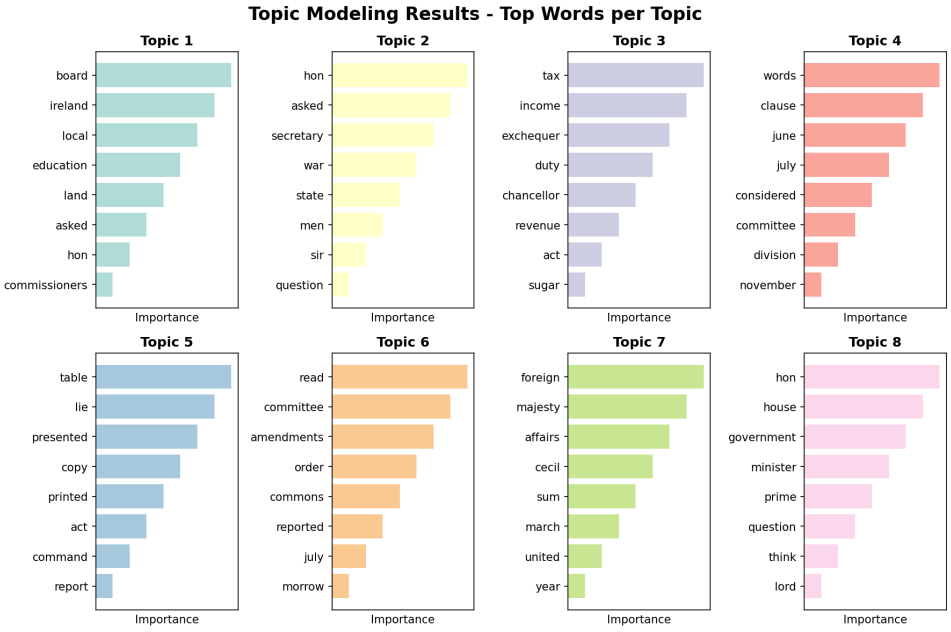
In addition, we do similar analyses for several key milestones in history. We look at 10 years before and after this time period, and in the case where the event lasted multiple years, we separate our analysis into pre/during/post the event under consideration.

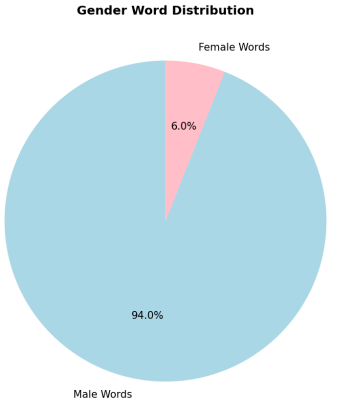
1918 - Women in Britain get partial suffrage and women are allowed to be MPs (1918 Representation of the People Act).

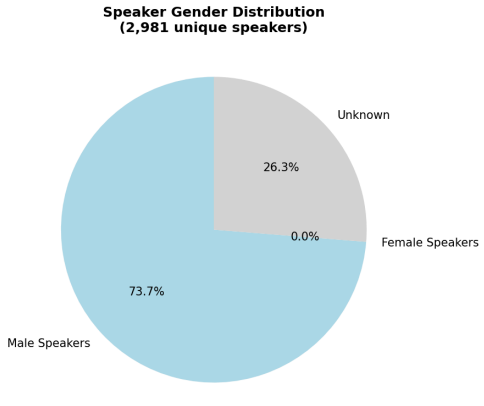
Pre-1918



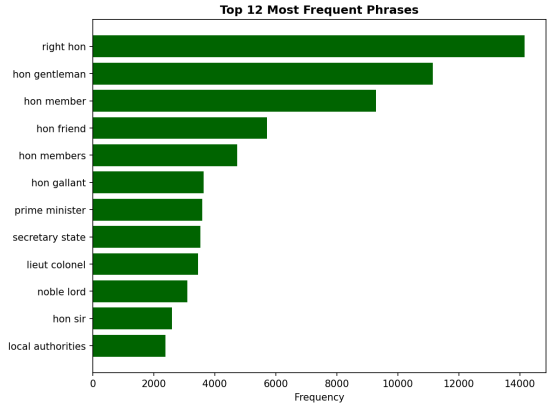


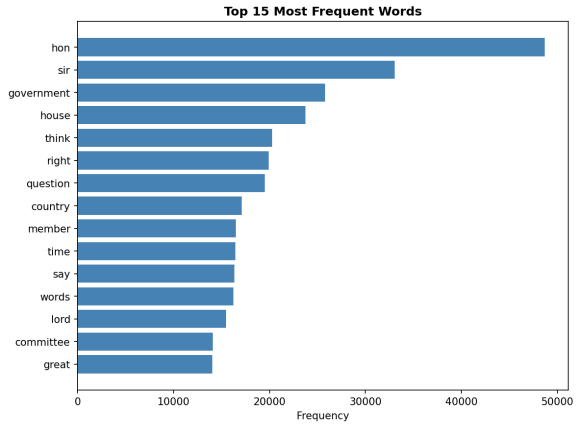


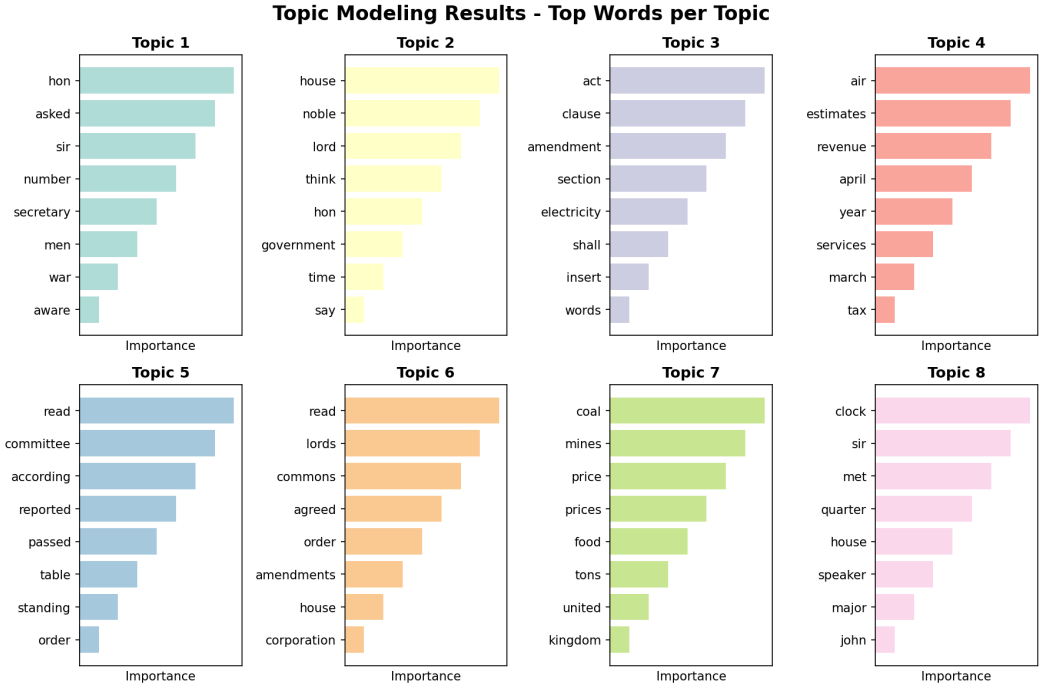


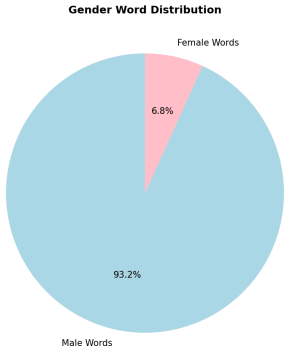


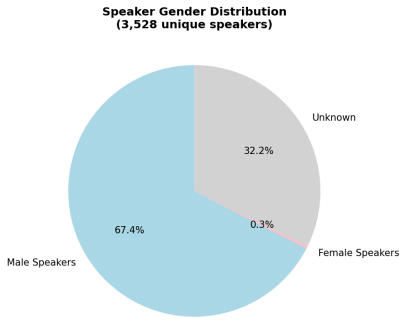
Post-1918


Top 12 Most Frequent Phrases
right hon
hon gentleman
hon member
hon friend
hon members
hon gallant
prime minister
secretary state
lieut colonel
noble lord
hon sir
local authorities
0
2000
4000
6000
8000
10000
12000
14000
Frequency


Top 15 Most Frequent Words
hon
sir
government
house
think
right
question
country
member
time
say
words
lord
committee
great
0
10000
20000
30000
40000
50000
Frequency

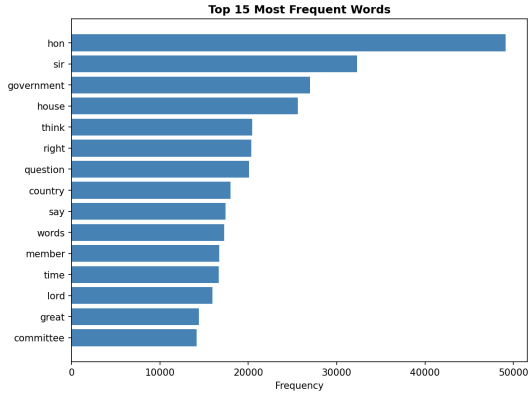

Topic Modeling Results - Top Words per Topic
Topic 1
hon
asked
sir
number
secretary
men
war
aware
Topic 2
house
noble
lord
think
hon
government
time
say
Topic 3
act
clause
amendment
section
electricity
shall
insert
words
Topic 4
air
estimates
revenue
april
year
services
march
tax
Topic 5
read
committee
according
reported
passed
table
standing
order
Topic 6
read
lords
commons
agreed
order
amendments
house
corporation
Topic 7
coal
mines
price
prices
food
tons
united
kingdom
Topic 8
clock
sir
met
quarter
house
speaker
major
john
Importance

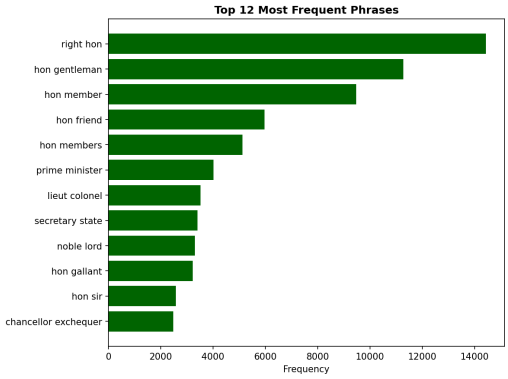

Gender Word Distribution
Female Words
6.8%
93.2%
Male Words

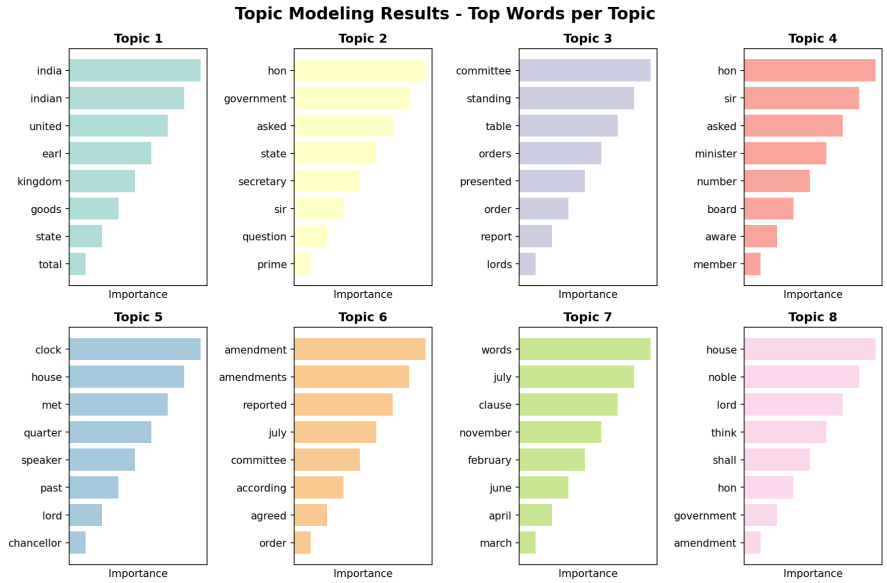

Speaker Gender Distribution
(3,528 unique speakers)
Unknown
32.2%
0.3%
67.4%
Female Speakers
Male Speakers

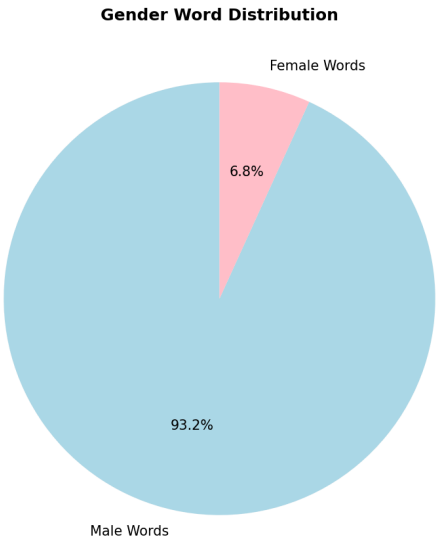
1928 - Women get universal suffrage (1928 Equal Franchise Act)

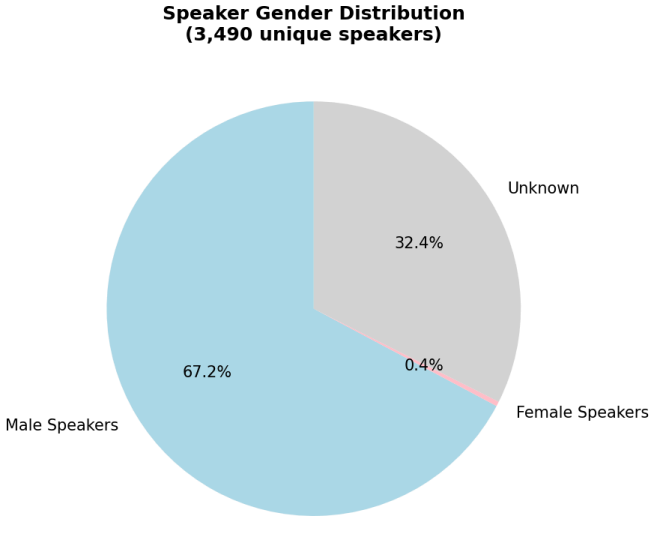
Pre-1928



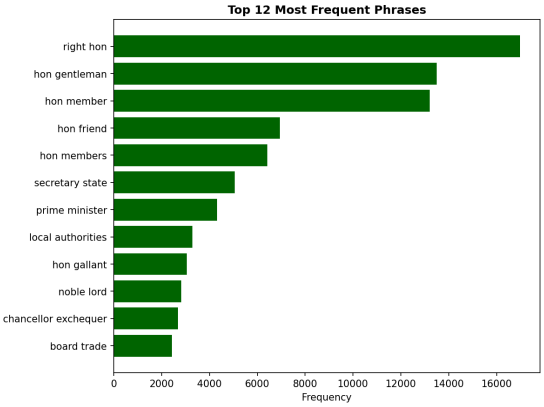


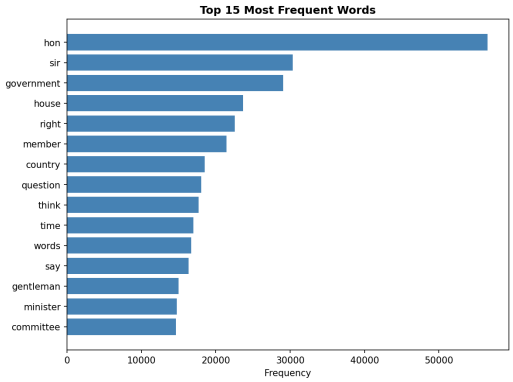


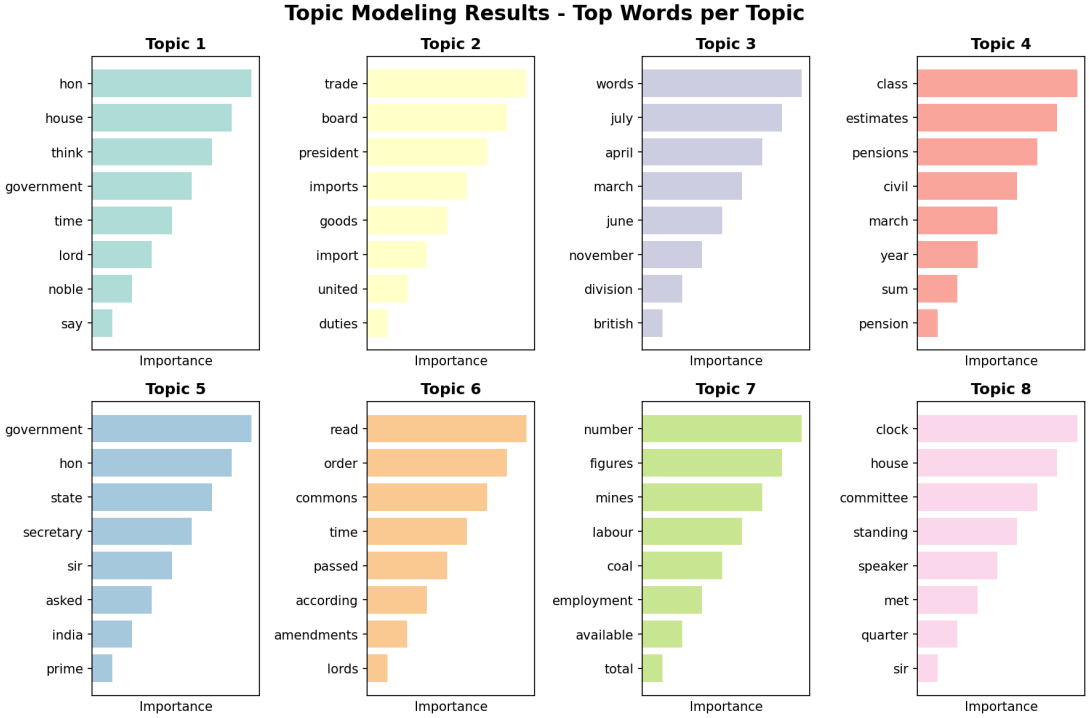


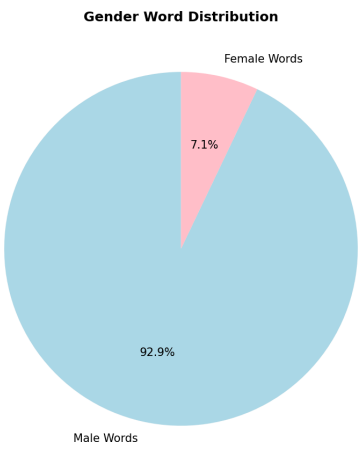


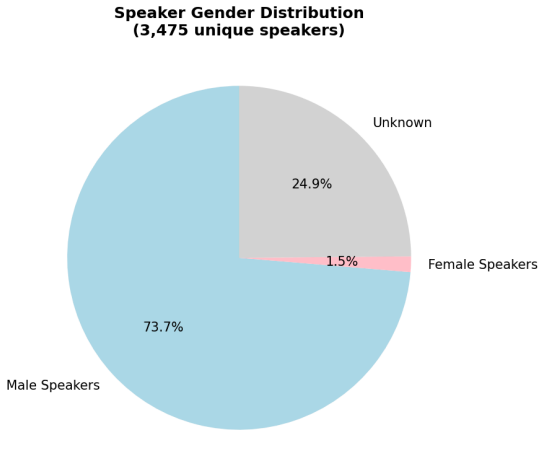
post-1928


Top 12 Most Frequent Phrases
right hon
hon gentleman
hon member
hon friend
hon members
secretary state
prime minister
local authorities
hon gallant
noble lord
chancellor exchequer
board trade
0
2000
4000
6000
8000
10000
12000
14000
16000
Frequency


Top 15 Most Frequent Words
hon
sir
government
house
right
member
country
question
think
time
words
say
gentleman
minister
committee
0
10000
20000
30000
40000
50000
Frequency

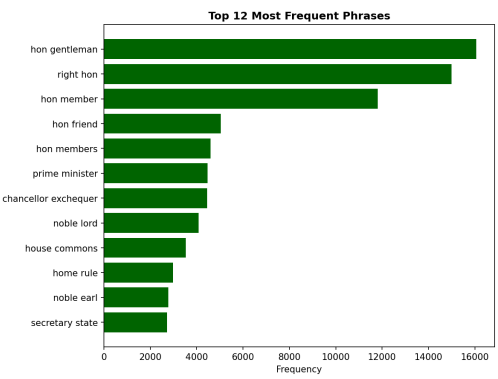

Topic Modeling Results - Top Words per Topic
Topic 1
hon
house
think
government
time
lord
noble
say
Topic 2
trade
board
president
imports
goods
import
united
duties
Topic 3
words
july
april
march
june
november
division
british
Topic 4
class
estimates
pensions
civil
march
year
sum
pension
Topic 5
government
hon
state
secretary
sir
asked
india
prime
Topic 6
read
order
commons
time
passed
according
amendments
lords
Topic 7
number
figures
mines
labour
coal
employment
available
total
Topic 8
clock
house
committee
standing
speaker
met
quarter
sir
Importance

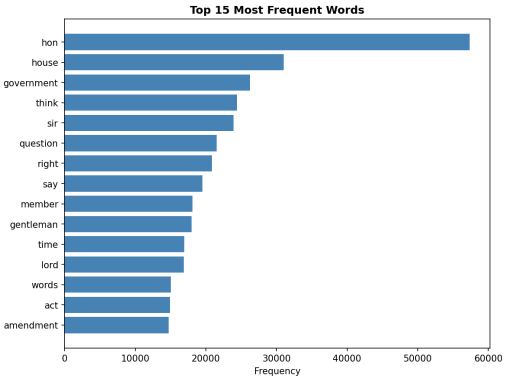

Gender Word Distribution
Female Words
7.1%
92.9%
Male Words

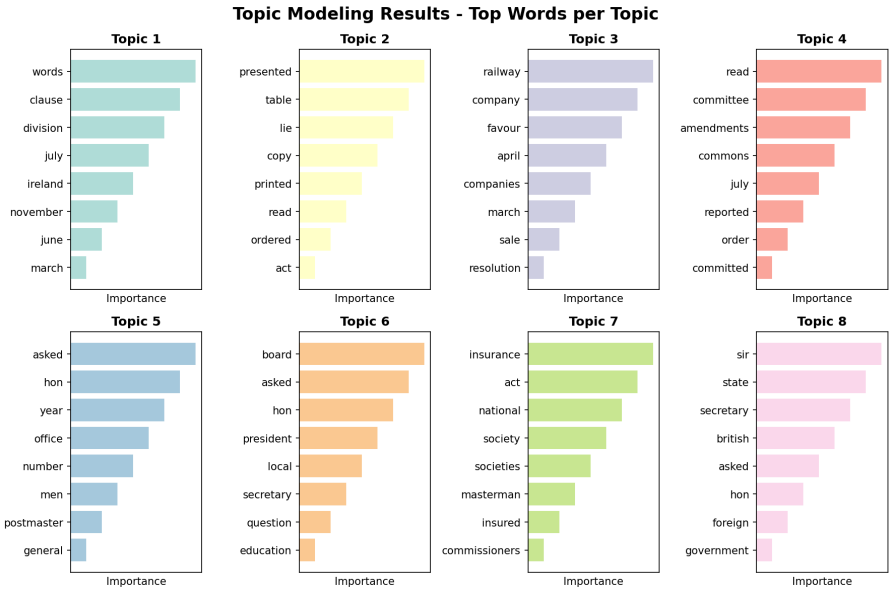

Speaker Gender Distribution
(3,475 unique speakers)
Unknown
24.9%
1.5%
Female Speakers
73.7%
Male Speakers

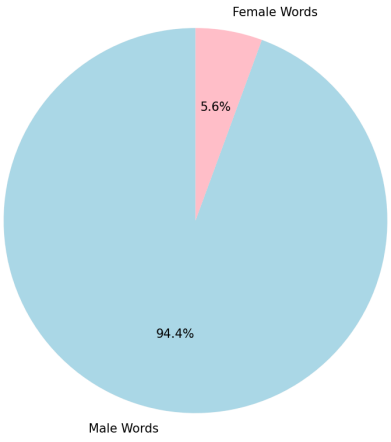
WW1 - 1914 - 1918

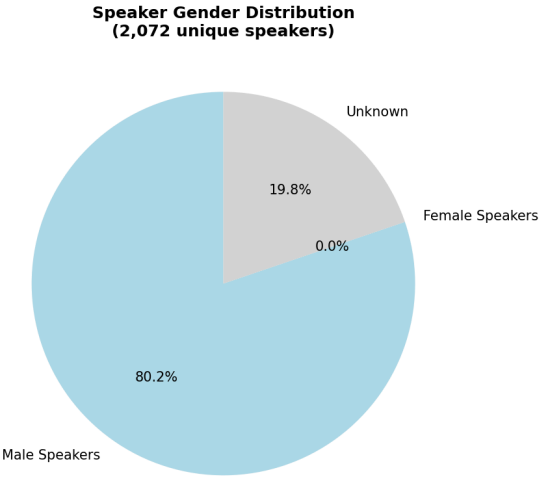
Pre-WW1



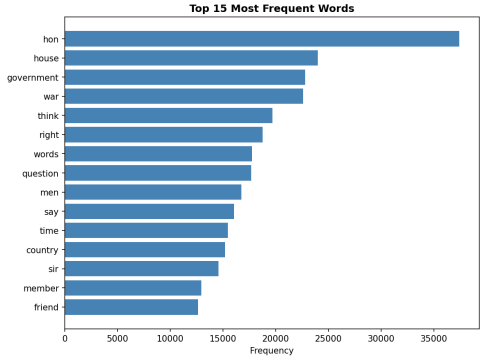


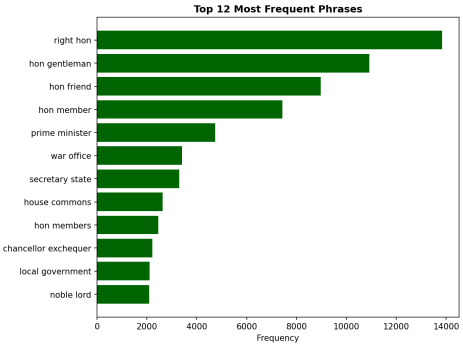


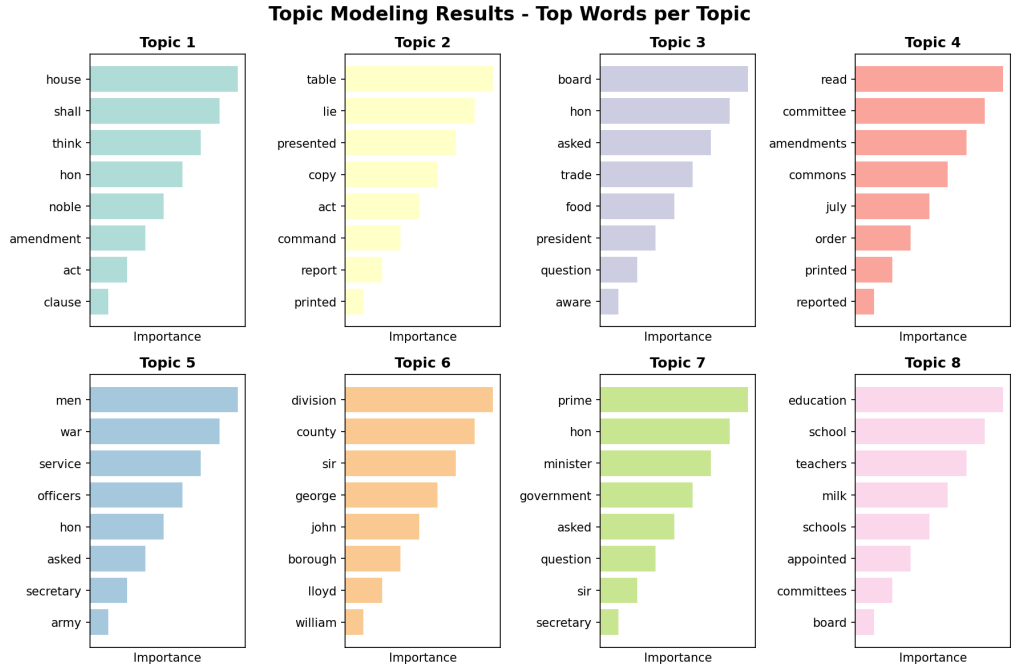


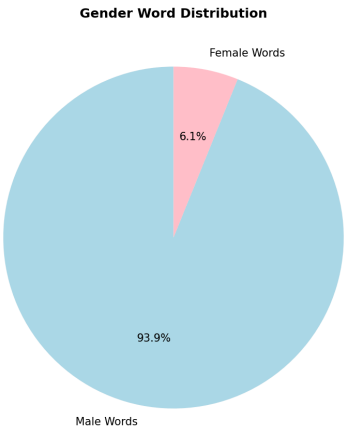


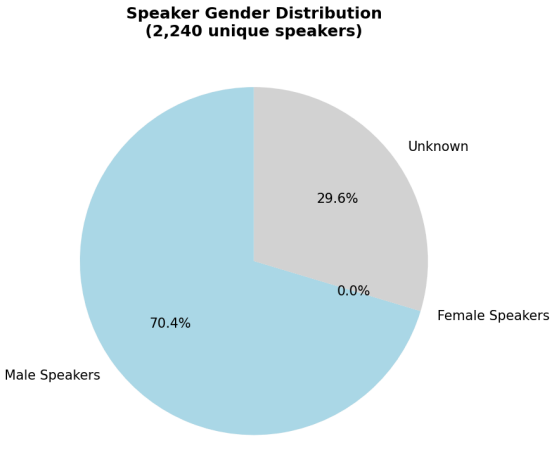
During-WW1


Top 15 Most Frequent Words
hon
house
government
war
think
right
words
question
men
say
time
country
sir
member
friend
0
5000
10000
15000
20000
25000
30000
35000
Frequency

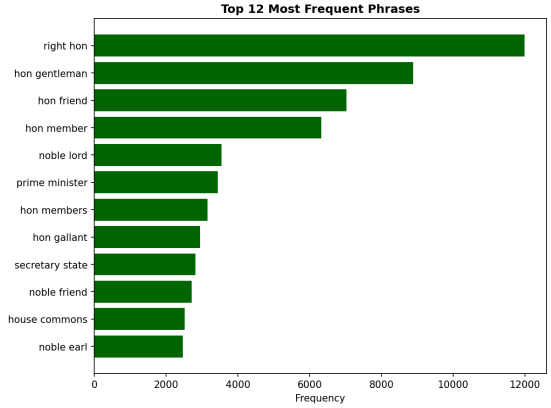

Top 12 Most Frequent Phrases
right hon
hon gentleman
hon friend
hon member
prime minister
war office
secretary state
house commons
hon members
chancellor exchequer
local government
noble lord
0
2000
4000
6000
8000
10000
12000
14000
Frequency

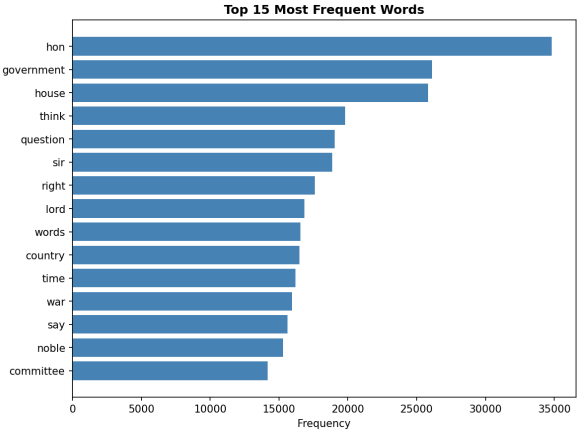

Topic Modeling Results - Top Words per Topic
Topic 1
house
shall
think
hon
noble
amendment
act
clause
Topic 2
table
lie
presented
copy
act
command
report
printed
Topic 3
board
hon
asked
trade
food
president
question
aware
Topic 4
read
committee
amendments
commons
july
order
printed
reported
Topic 5
men
war
service
officers
hon
asked
secretary
army
Topic 6
division
county
sir
george
john
borough
lloyd
william
Topic 7
prime
hon
minister
government
asked
question
sir
secretary
Topic 8
education
school
teachers
milk
schools
appointed
committees
board
Importance

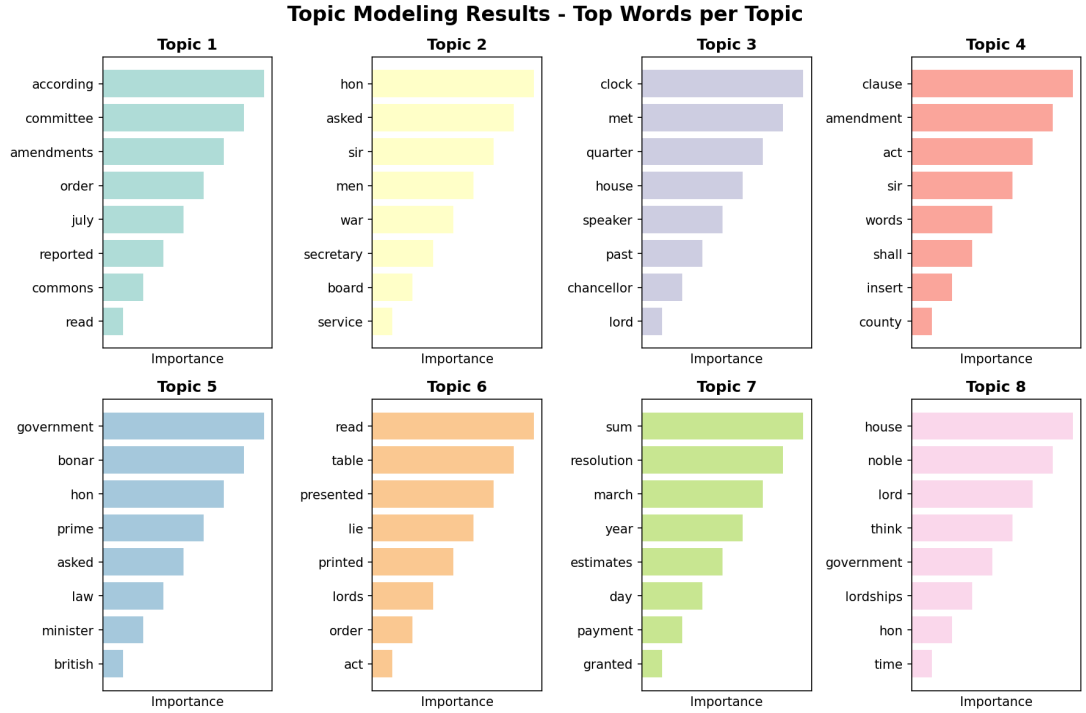

Gender Word Distribution
Female Words
6.1%
93.9%
Male Words

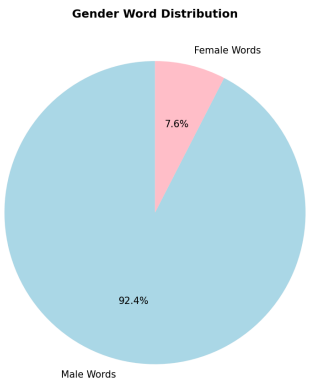

Speaker Gender Distribution
(2,240 unique speakers)
Unknown
29.6%
0.0%
Female Speakers
70.4%
Male Speakers

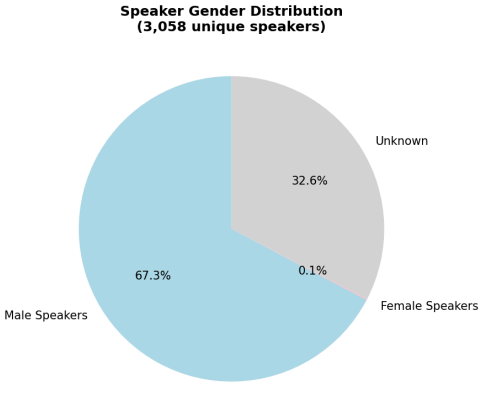
Post-WW1


Top 12 Most Frequent Phrases
right hon
hon gentleman
hon friend
hon member
noble lord
prime minister
hon members
hon gallant
secretary state
noble friend
house commons
noble earl
0
2000
4000
6000
8000
10000
12000
Frequency


Top 15 Most Frequent Words
hon
government
house
think
question
sir
right
lord
words
country
time
war
say
noble
committee
0
5000
10000
15000
20000
25000
30000
35000
Frequency

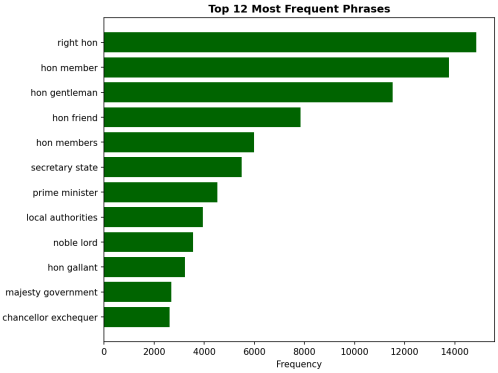

Topic Modeling Results - Top Words per Topic
Topic 1
according
committee
amendments
order
july
reported
commons
read
Importance
Topic 2
hon
asked
sir
men
war
secretary
board
service
Importance
Topic 3
clock
met
quarter
house
speaker
past
chancellor
lord
Importance
Topic 4
clause
amendment
act
sir
words
shall
insert
county
Importance
Topic 5
government
bonar
hon
prime
asked
law
minister
british
Importance
Topic 6
read
table
presented
lie
printed
lords
order
act
Importance
Topic 7
sum
resolution
march
year
estimates
day
payment
granted
Importance
Topic 8
house
noble
lord
think
government
lordships
hon
time
Importance

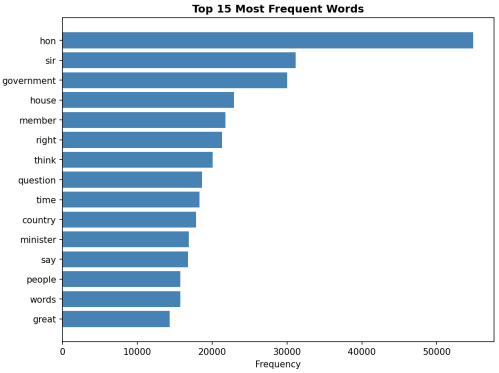

Gender Word Distribution
Female Words
7.6%
92.4%
Male Words

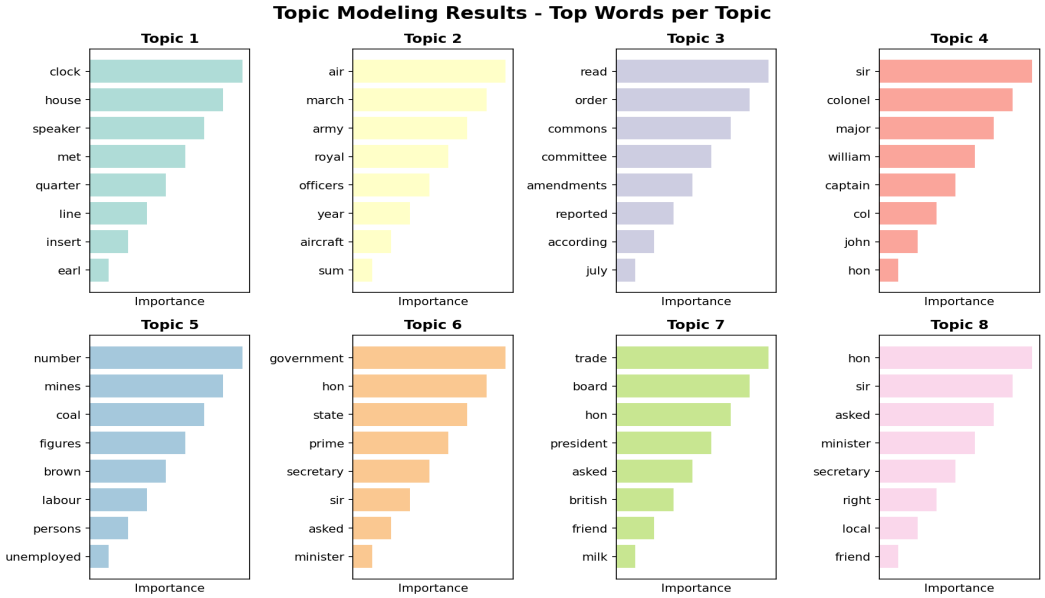

Speaker Gender Distribution
(3,058 unique speakers)
Unknown
32.6%
0.1%
67.3%
Male Speakers
Female Speakers

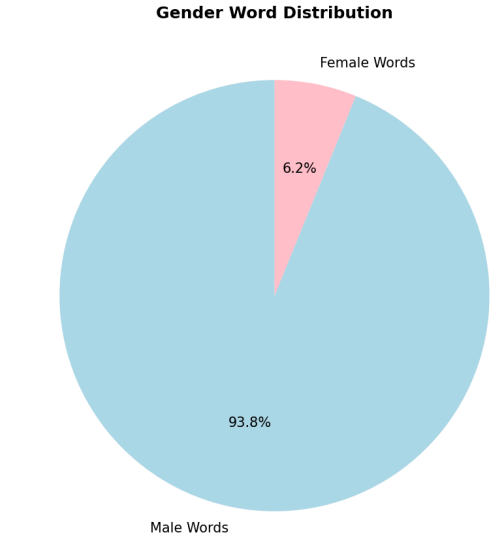
WW2 - 1939-1945

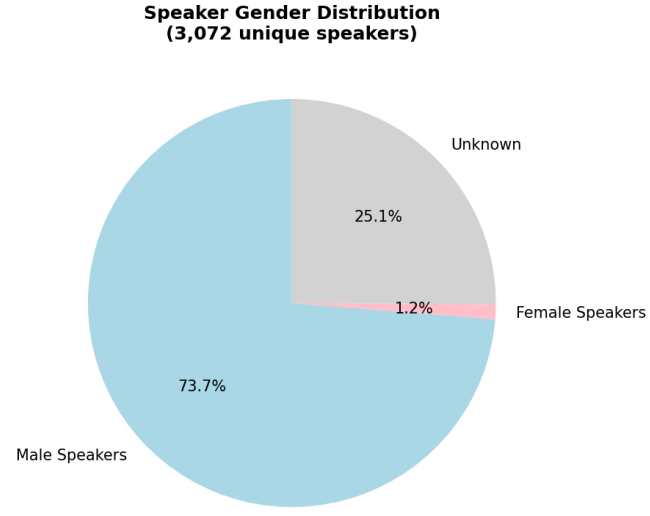
Pre-WW2



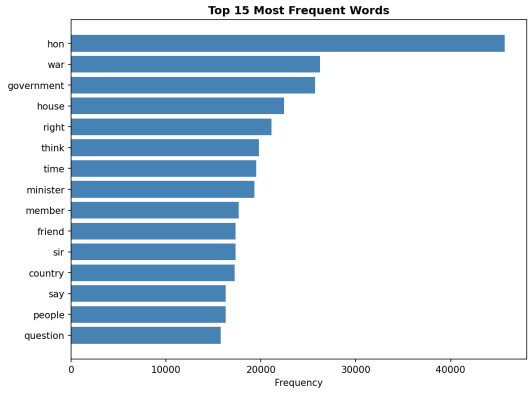


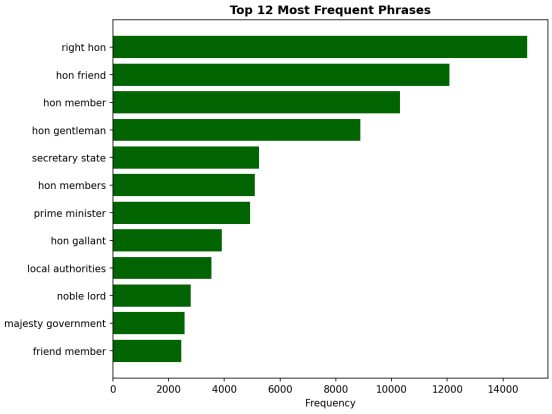


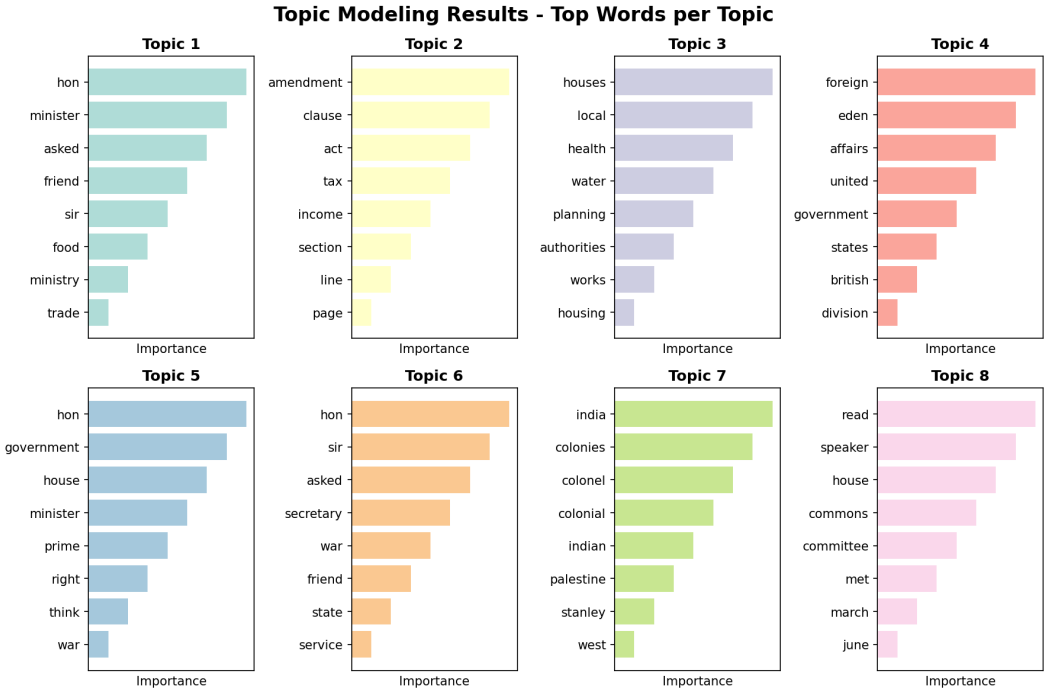


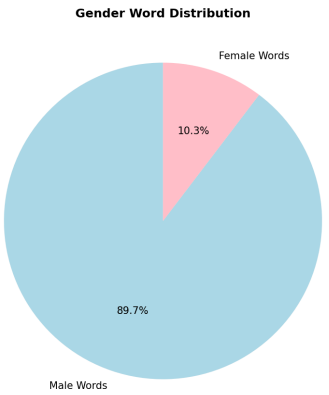


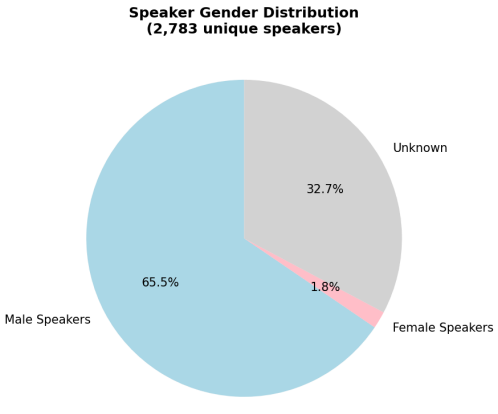
During-WW2


Top 15 Most Frequent Words
hon
war
government
house
right
think
time
minister
member
friend
sir
country
say
people
question
0
10000
20000
30000
40000
Frequency

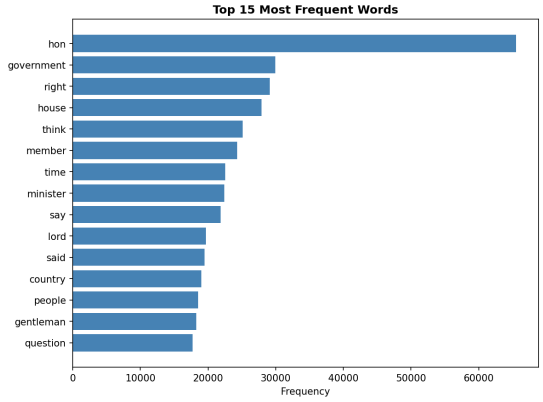

Top 12 Most Frequent Phrases
right hon
hon friend
hon member
hon gentleman
secretary state
hon members
prime minister
hon gallant
local authorities
noble lord
majesty government
friend member
0
2000
4000
6000
8000
10000
12000
14000
Frequency

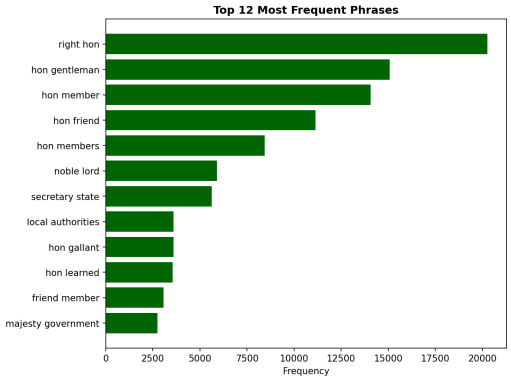

Topic Modeling Results - Top Words per Topic
Topic 1
hon
minister
asked
friend
sir
food
ministry
trade
Topic 2
amendment
clause
act
tax
income
section
line
page
Topic 3
houses
local
health
water
planning
authorities
works
housing
Topic 4
foreign
eden
affairs
united
government
states
british
division
Topic 5
hon
government
house
minister
prime
right
think
war
Topic 6
hon
sir
asked
secretary
war
friend
state
service
Topic 7
india
colonies
colonel
colonial
indian
palestine
stanley
west
Topic 8
read
speaker
house
commons
committee
met
march
june
Importance

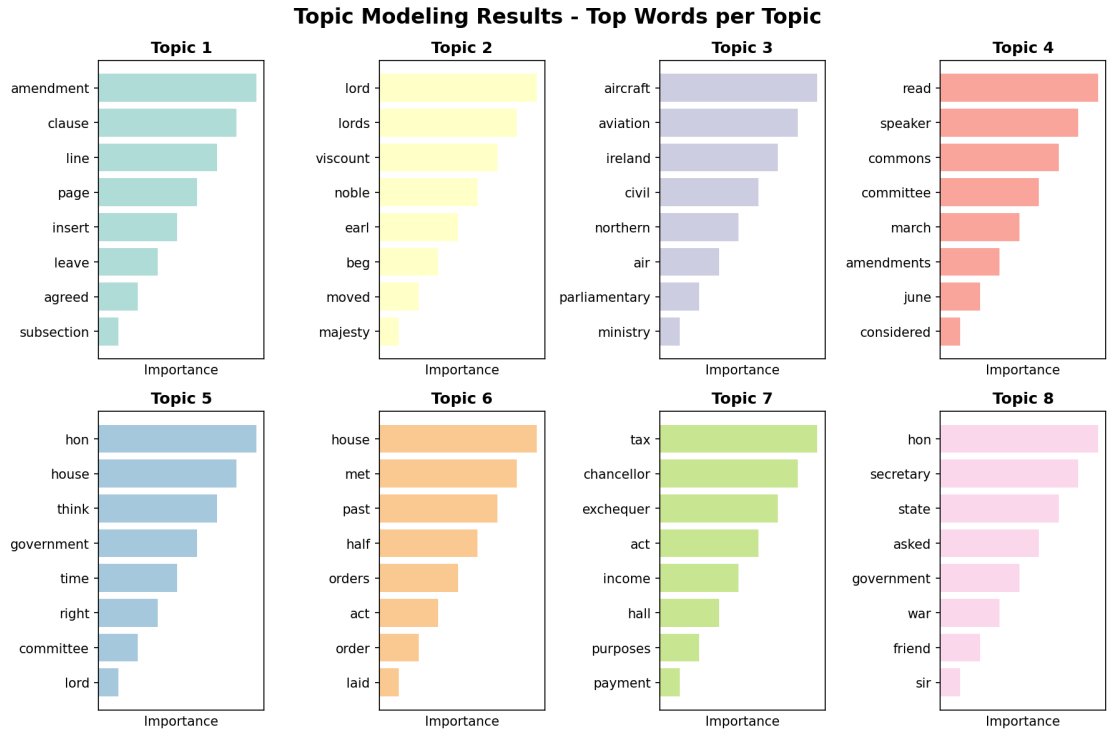

Gender Word Distribution
Female Words
10.3%
89.7%
Male Words

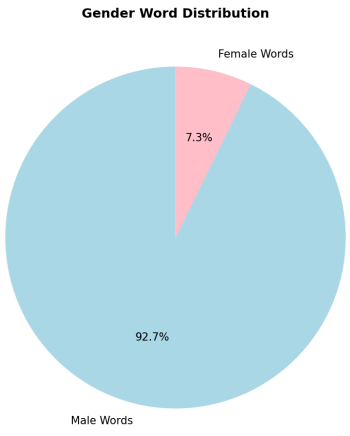

Speaker Gender Distribution
(2,783 unique speakers)
Unknown
32.7%
65.5%
1.8%
Male Speakers
Female Speakers

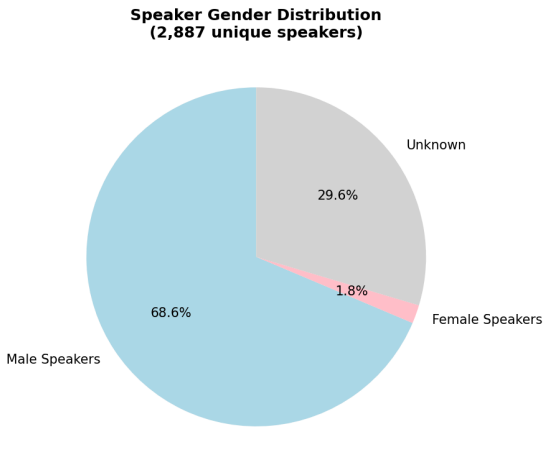
Post-WW2


Top 15 Most Frequent Words
hon
government
right
house
think
member
time
minister
say
lord
said
country
people
gentleman
question
0
10000
20000
30000
40000
50000
60000
Frequency


Top 12 Most Frequent Phrases
right hon
hon gentleman
hon member
hon friend
hon members
noble lord
secretary state
local authorities
hon gallant
hon learned
friend member
majesty government
0
2500
5000
7500
10000
12500
15000
17500
20000
Frequency

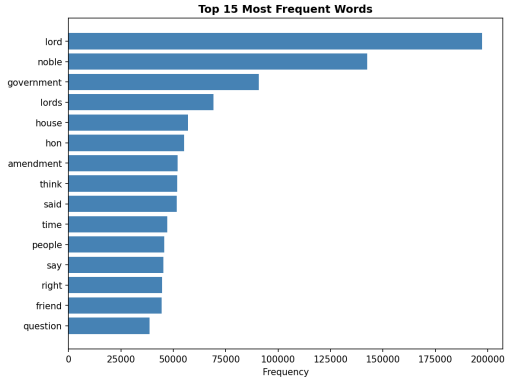

Topic Modeling Results - Top Words per Topic
Topic 1
amendment
clause
line
page
insert
leave
agreed
subsection
Topic 2
lord
lords
viscount
noble
earl
beg
moved
majesty
Topic 3
aircraft
aviation
ireland
civil
northern
air
parliamentary
ministry
Topic 4
read
speaker
commons
committee
march
amendments
june
considered
Topic 5
hon
house
think
government
time
right
committee
lord
Topic 6
house
met
past
half
orders
act
order
laid
Topic 7
tax
chancellor
exchequer
act
income
hall
purposes
payment
Topic 8
hon
secretary
state
asked
government
war
friend
sir
Importance

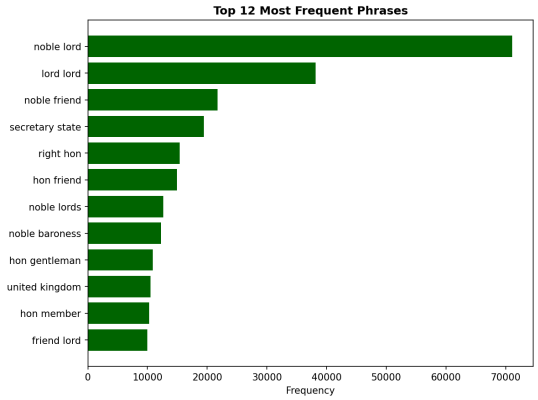

Gender Word Distribution
Female Words
7.3%
92.7%
Male Words

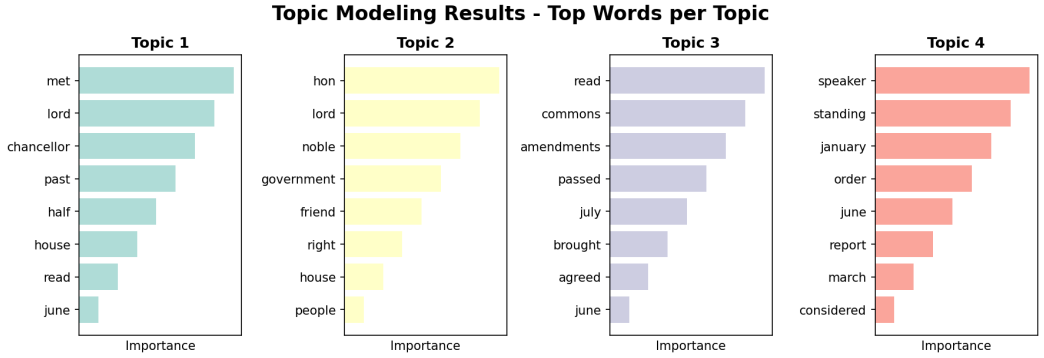

Speaker Gender Distribution
(2,887 unique speakers)
Unknown
29.6%
1.8%
Female Speakers
68.6%
Male Speakers

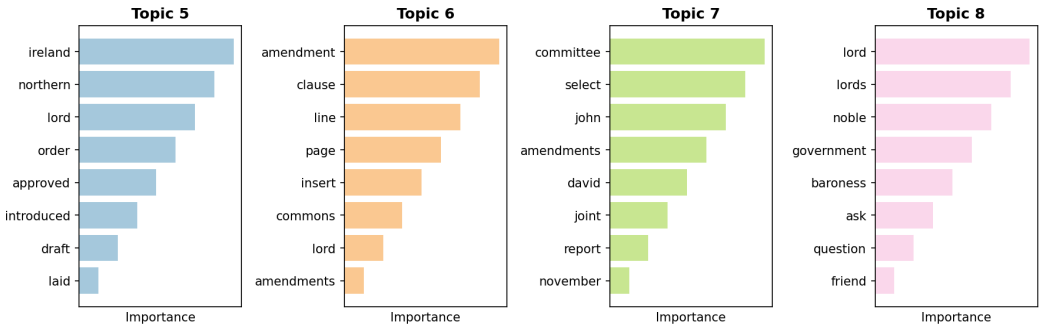
1979 - 1990 Margaret Thatcher is PM (first woman PM in the UK)

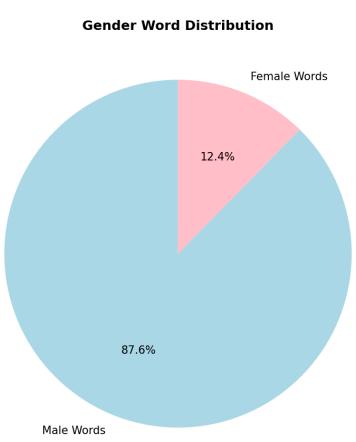
Pre-Thatcher:

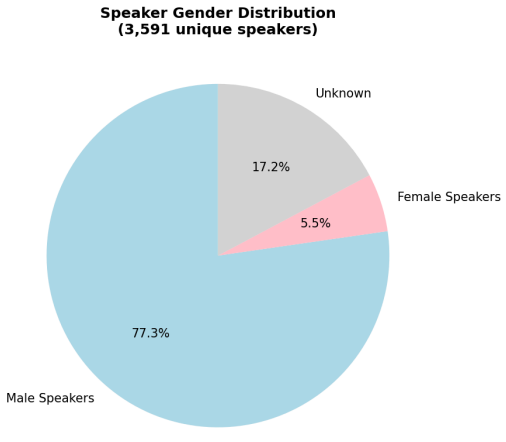




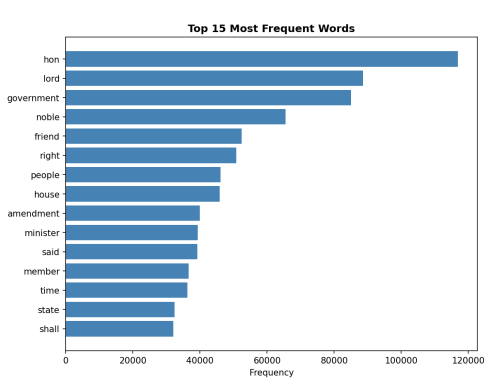


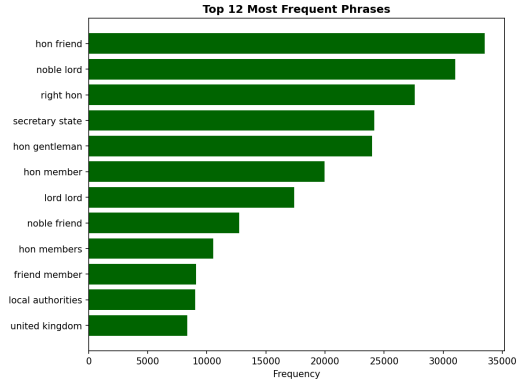


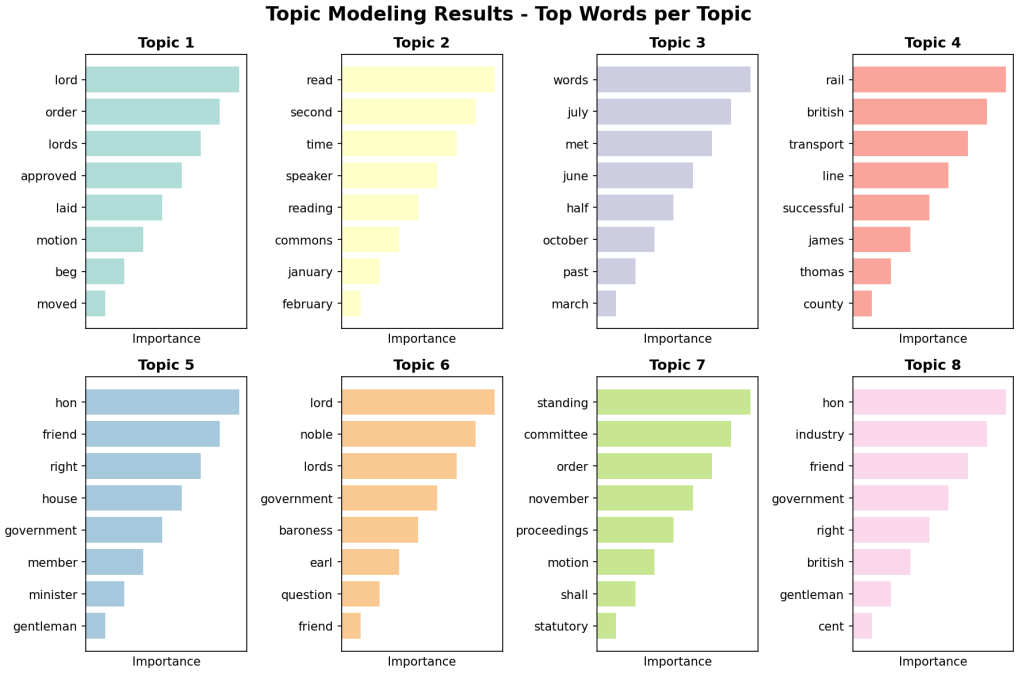


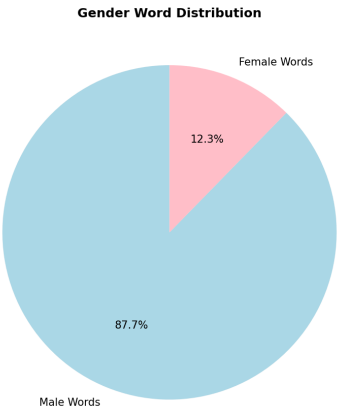


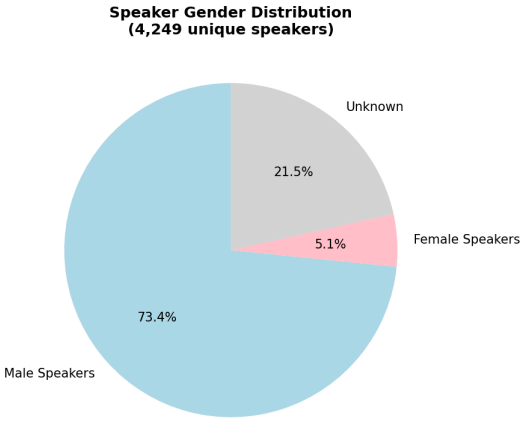
During-Thatcher



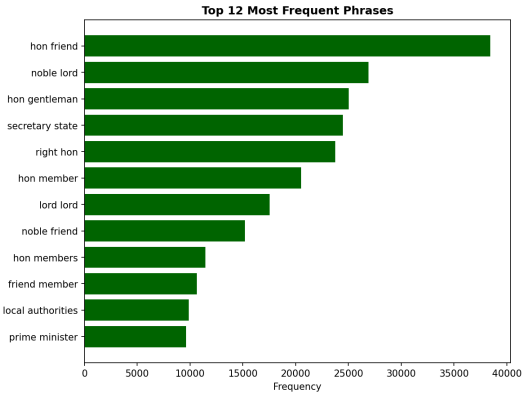


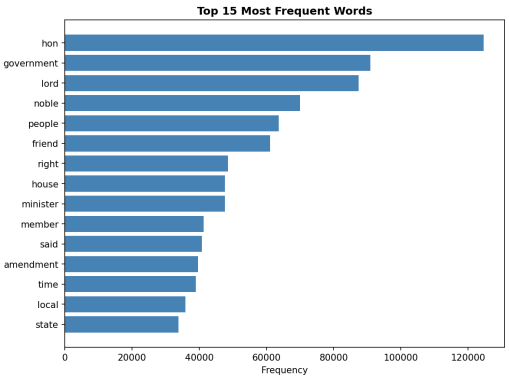


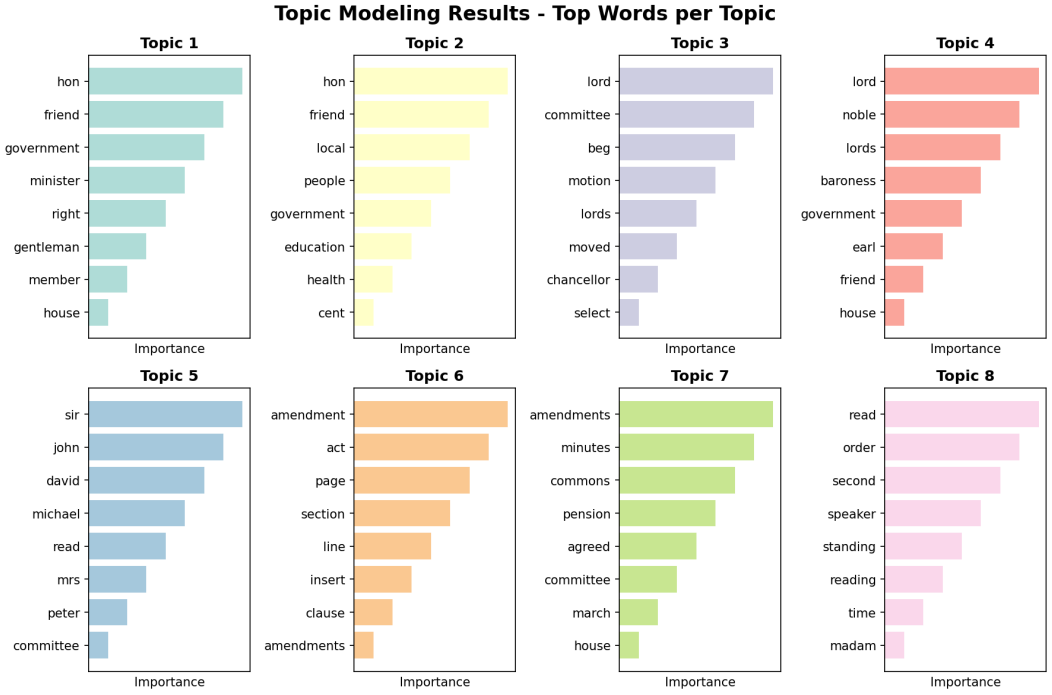


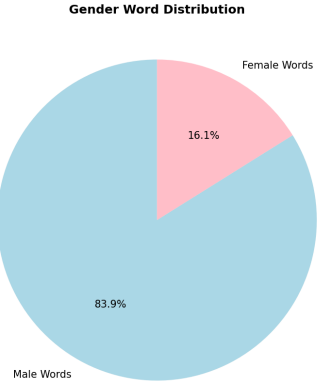


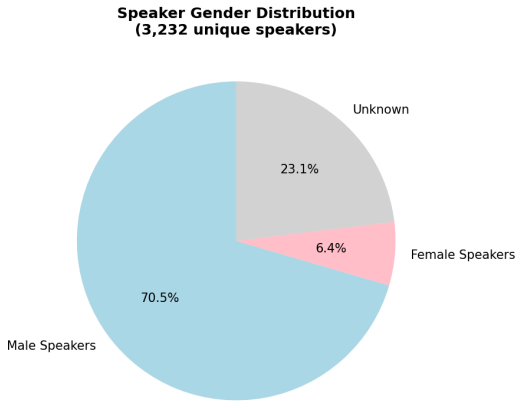
Post-Thatcher











With the analysis we've completed for these different historical periods, we can see that there's a lot of commonalities in bigram/unigram analysis, as well as topics found. We seem to be mainly extracting parliamentary procedures and specific terminology, rather than the substance of the debates. Therefore, our next steps would be:

1. Develop a method to extract debate content versus parliamentary terminology.
2. Current heuristic based method for extracting MP gender is likely somewhat inaccurate, let's find a list of all members of houses in this period and cross-reference against other sources (such as wikipedia) to get a more accurate gender count
3. Develop a DSPy pipeline (potentially, still need to look into this) to use an LLM to extract and clean up debate text.
4. Improve speed of analysis so that we can run it locally