3.1

The purpose of this section is to describe and explain the procedures utilized in my study on the frequency of modals and semi-modals in the THC and COHA corpora. Enough detail has been supplied to explain how the procedures were carried out.

Note the URL to access the COHA corpus. The level of access I needed required the creation of a free account. Anyone can apply for, and receive permission to, open an account.

Next, I wanted to collect frequency data for both modals and semi-modals on a historical basis. For this aspect of the search to work, the bottom-left box (Sorting and Limits) has to have Frequency highlighted under Sorting, and the Frequency needs to be set to 10 years. The next graphic shows how to set up these parameters for COHA. Again, this process can be replicated by anyone wishing to verify my analyses.
Now let's examine the result:

The graphic above can't be seen well, because it is horizontally ruled, so let's zoom in:
The key point here is to note that, for each word in the search, the automated results return the frequency of that word for every 10 years in COHA. For example, in the search for *must*, the graphic above shows that there were 2198 uses of this word in 1810, 8589 in 1820, and 15693 in 1830. In fact, given the parameters I entered earlier, COHA returned a frequency for the word *must* in every decade from 1810 to 2000. Note that these results are for raw frequency. In order to ensure more valid comparisons, I delimited my results to words per million, which accounted for the different size of the virtual corpus in different years.

What these figures allowed me to do is to perform some statistical analyses on how the frequencies of *must* (and the other modals and semi-modals in my analysis) changed over time. So now these data can be entered in a program for statistical analysis. I used Stata for this purpose, but I used also eViews for one of the analyses.

1. Unicode is supported; see `help unicode_advice`.
2. Maximum number of variables is set to 5000; see `help set_maxvar`.
Note that, for the variable names in Stata and eViews, THC refers to the Hansard Corpus, while COHA is the Corpus of Historical American English. The names of the variables can be more easily understood once the prefixes are understood, but there are some special terms that require explanation. Note that I created average (semimodal_avg and modal_avg) variables by averaging out all of the modals and semi-modals in the THC and COHA corpora. This procedure allowed me to trace the frequency changes of all modals and semi-modals, which was an important consideration when trying to understand the higher-level trends in frequency changes. Of course, I analyzed the frequency change of individual modals and semi-modals as well, but the creation of a single, averaged variable for both modals and semi-modals let me achieve insights that couldn’t have been generated by analyses of the individual modals and semi-modals.
Note also that the ln prefix refers to a logarithmic transformation. Because raw frequencies were quite different (in terms of magnitude) between modals and semi-modals, I needed a method to compare them on the same graph. Using a log-transform procedure allowed me to achieve this kind of comparative insight, which was an important part of the analysis. Consider the graph below:

Also, note that the syntax to reproduce this graphic in Stata is as follows:

```
tsline coha_semimodal_avg coha_modal_avg
```

Note that the raw count of modal frequencies is much higher, so it isn’t really possible to see the convergence of modals and semi-modals. Now, when log-transformed data are used, the result is as follows:

Note that the syntax to reproduce this graphic in Stata is as follows:

```
tsline lncoha_semimodal_avg lncoha_modal_avg
```

What takes place in the log-transformed model is a much easier illustration of the convergence of the frequency of modals and semi-modals in COHA over time. This analysis tells us that, in COHA, the ratio of semi-modals to modals has been increased over time. Following Smith’s substitution thesis, I wanted to see if the relationship between modal decline and semi-modal increase in COHA could be quantified through linear regression. In Stata, the command for this analysis
would be `reg coha_modal_av coha_semimodal_avg`, and the results are as follows:

```
 . reg coha_modal_avg coha_semimodal_avg
 Source |      SS     df  MS  Number of obs = 20
        |           |           |          |
 Model  | 504309.655 |    1 | 504309.655 | Prob > F = 0.0000
 Residual | 197736.128 |   18 | 10985.3404 | R-squared = 0.7183
        |           |           |          |
 Total  | 702045.783 |   19 | 36949.778  | Adj R-squared = 0.7027
        |           |           |          |
                 |        |       | Root MSE = 104.81
```

These results were actually part of my study; a Ctrl-F search for 45.91 will identify where they occur in the document. I created several histograms in Stata for inclusion in the thesis. The dialog box below demonstrates how histograms are created in Stata.

I used eViews for modeling the effect (as an impulse response function, or IRF) of large changes in semi-modals on the subsequent frequencies of modals. I performed this analysis for both TCH and COHA. Note that, for the IRF to work in eViews, the 2 variables of interest (in this case, modal and semi-modals averages for COHA for every decade from 1810 to 2000) have to be opened as part of a vector autoregression, or VAR:
Once the variables are opened as a VAR, then the IRF can be generated as follows:
In this dialog box, semimodal average is the independent (impulse) variable, and modsl average is the response (dependent) variable. Once OK is clicked, the model generates the graphs I used in my paper.

The same procedures that I used for COHA were used for THC as well. First, let me show how the Hansard Corpus can be retrieved:

![corpus.byu.edu](image_url)

<table>
<thead>
<tr>
<th>Corpus</th>
<th># words</th>
<th>language/dialect</th>
<th>time period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia Corpus (with virtual corpora)</td>
<td>1.9 billion</td>
<td>English</td>
<td>-2014</td>
</tr>
<tr>
<td>Global Web-Based English (GloWbE)</td>
<td>1.9 billion</td>
<td>20 countries</td>
<td>2012-13</td>
</tr>
<tr>
<td>Corpus of Contemporary American English (COCA)</td>
<td>520 million</td>
<td>American</td>
<td>1990-2015</td>
</tr>
<tr>
<td>Corpus of Historical American English (COHA)</td>
<td>400 million</td>
<td>American</td>
<td>1810-2009</td>
</tr>
<tr>
<td>TIME Magazine Corpus</td>
<td>100 million</td>
<td>American</td>
<td>1923-2006</td>
</tr>
<tr>
<td>Corpus of American Soap Operas</td>
<td>100 million</td>
<td>American</td>
<td>2001-2012</td>
</tr>
<tr>
<td>British National Corpus (BYU-BNC)</td>
<td>100 million</td>
<td>British</td>
<td>1980-1993</td>
</tr>
<tr>
<td>Strathy Corpus (Canada)</td>
<td>50 million</td>
<td>Canadian</td>
<td>1970s-2000s</td>
</tr>
<tr>
<td>Hansard Corpus (British Parliament)</td>
<td>1.6 billion</td>
<td>British</td>
<td>1803-2005</td>
</tr>
</tbody>
</table>
BYU has a corpus site where both COHA and TCH can be accessed. Note that, in order to pull up modal or semi-modal frequencies by decade, the same procedure is followed for THC as was followed for COHA (see screen shot on following page). Again, the key with this form of analysis is to choose 10 for the Frequency, and to highlight the Frequency box. Once these steps are taken, then the resulting readout sorts the search word into frequencies per decade. Now that, as with COHA, I configured the THC results to be frequency per million words, which allowed better comparisons from year to year (which is an important consideration for all virtual corpora, given that the size of each corpus changes from year to year). Raw frequencies are therefore not an appropriate input for analysis.

Finally, here are partial screenshots of my Stata dataset:
Note that all of the raw data for the study have been presented in full in the body of the study itself.

### Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).