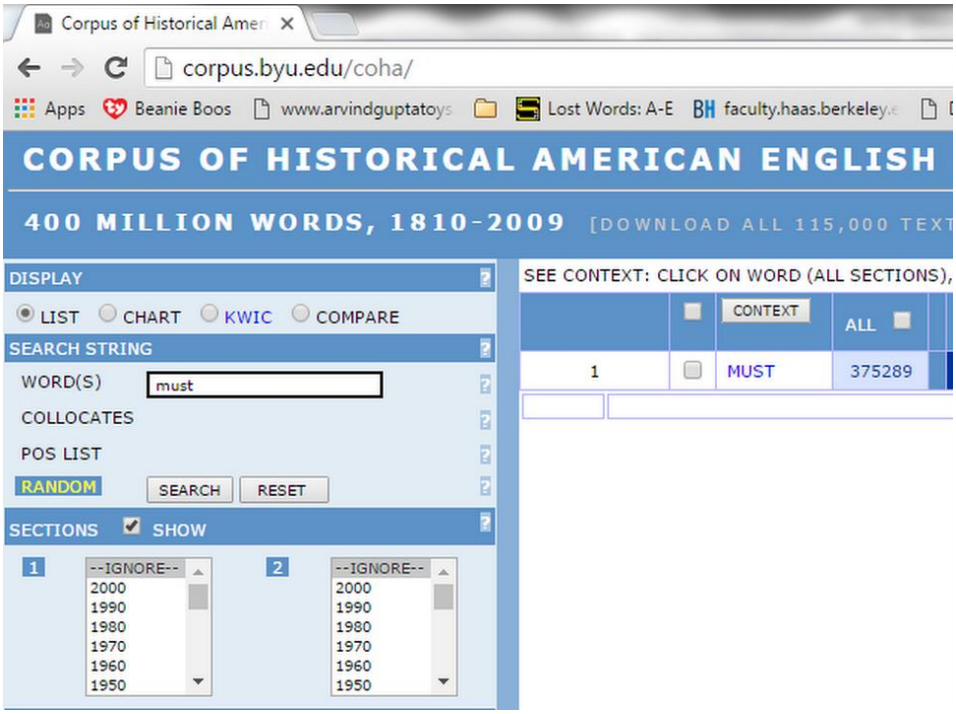
International Journal of English and Cultural Studies
Vol. 1, No. 1; May 2018
ISSN 2575-811X E-ISSN 2575-8101
Published by Redfame Publishing
URL: http://ijecs.redfame.com

Supplementary Document

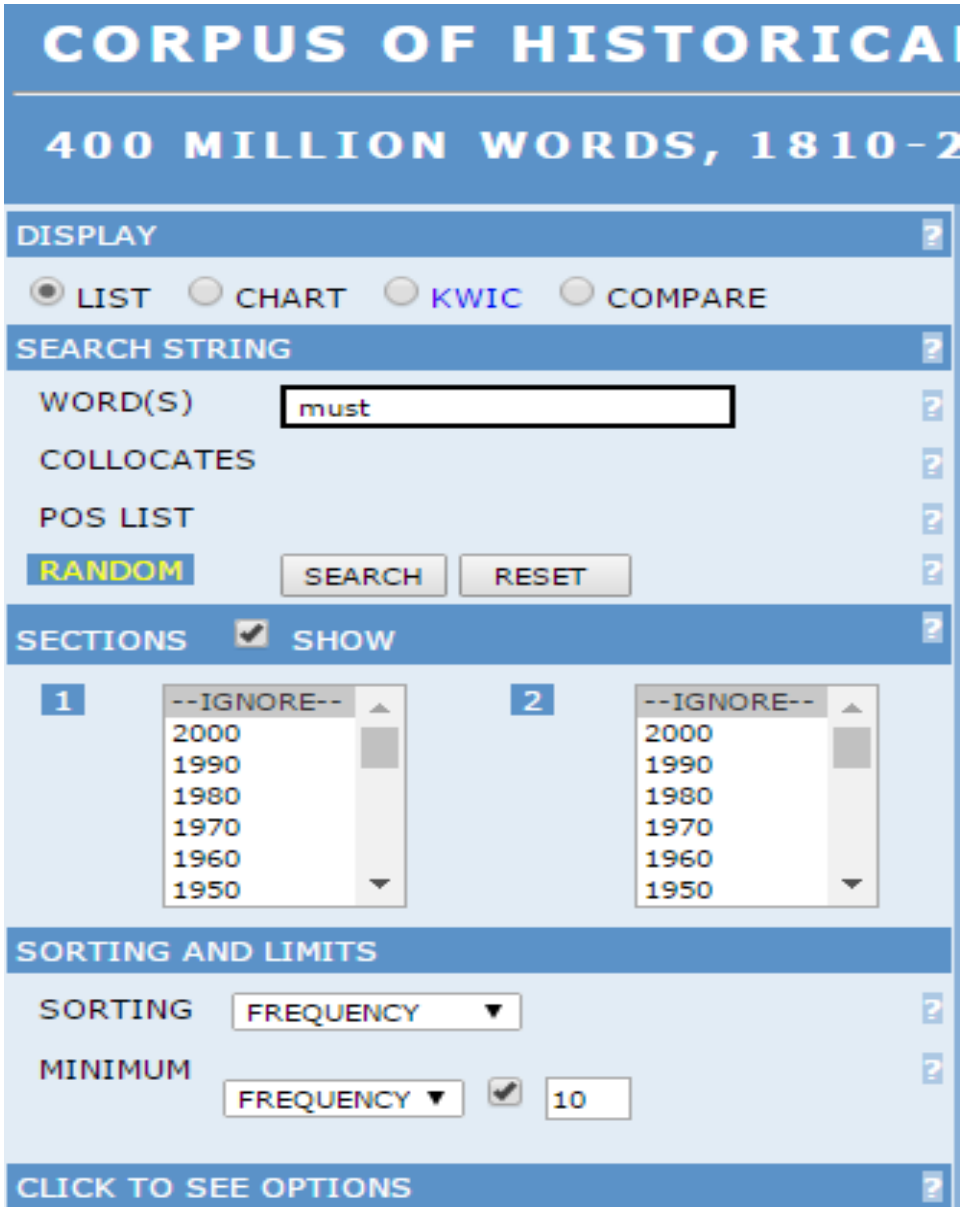
3.1

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.



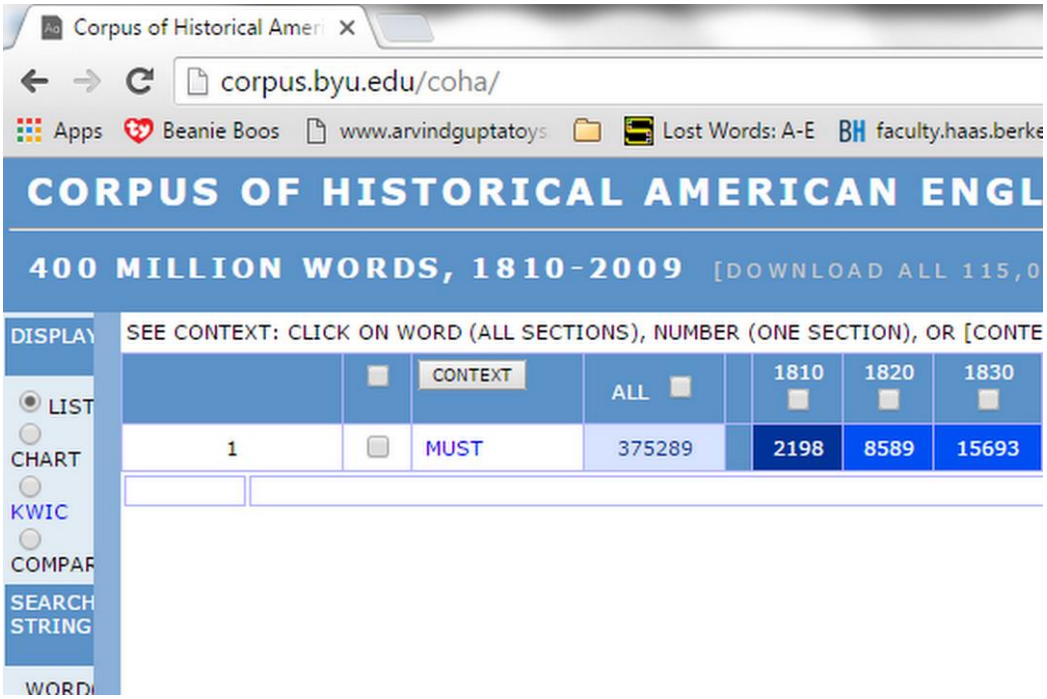
Note the URL to access the COHA corpus. The level of access I needed required the creation of a free acount. Anyone can apply for, and receive permission to, open an account.

Next, I wanted to collecy 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.

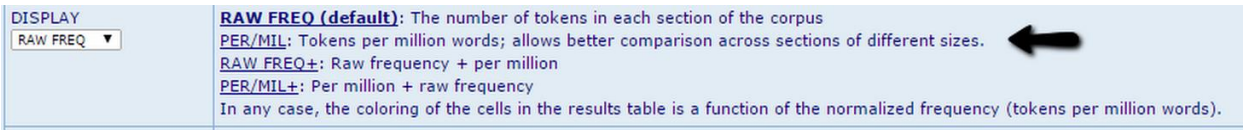


Now let's examine the result:

The graphic above can't be seen well, because it is horizontally ruled, so let's zoom in:

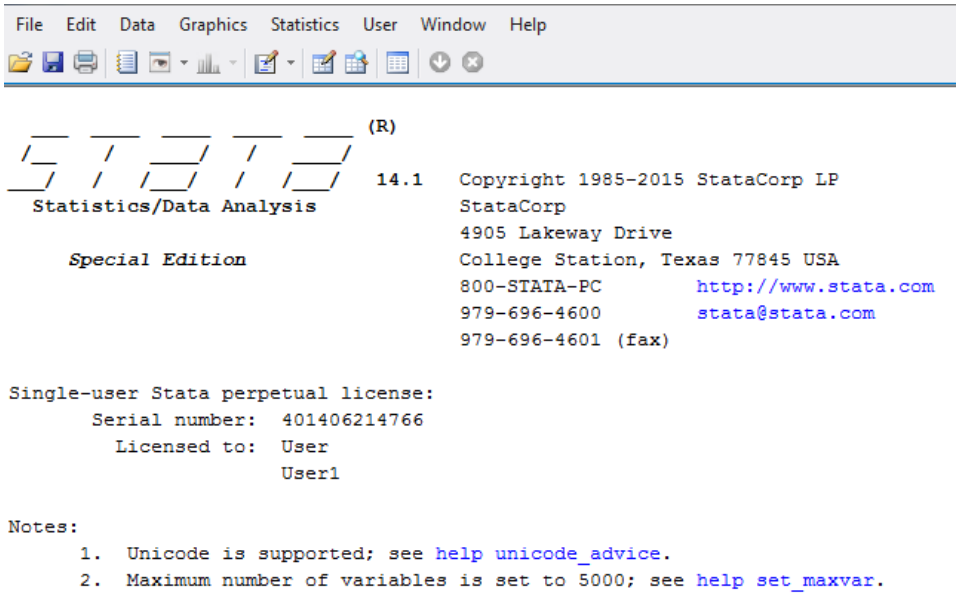


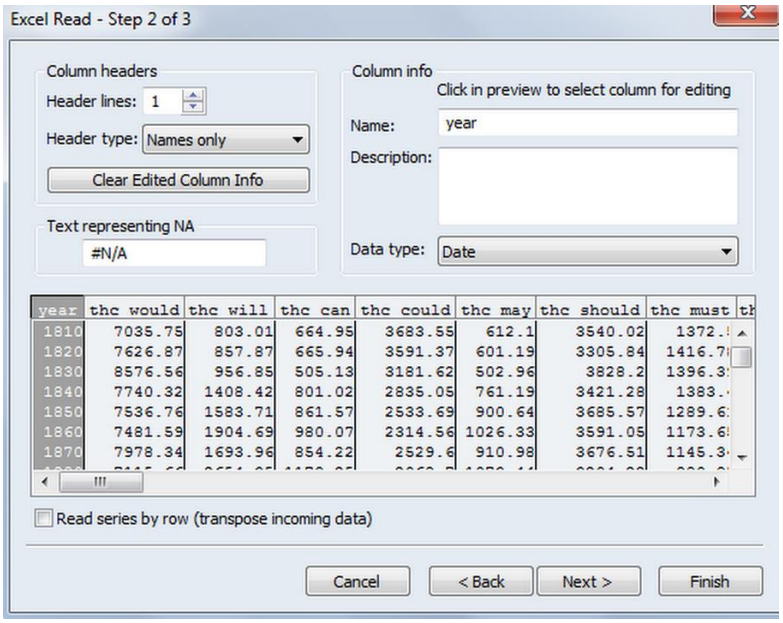
The key point here is to note that, for each word in the search, the automated results return the frequency of that word for every 10 years in COHA. For example, in the search for *must*, the graphic above shows that there were 2198 uses of this word in 1810, 8589 in 1820, and 15693 in 1830. In fact, 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.

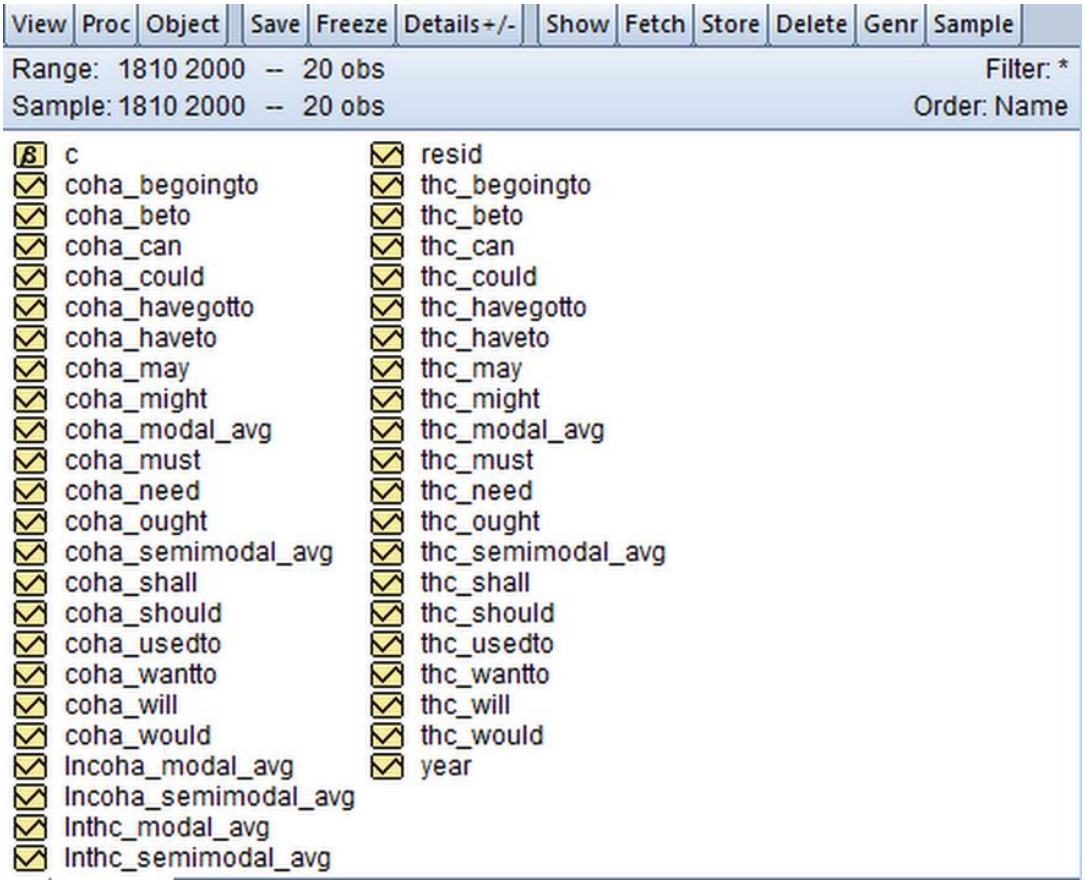


What these figures allowed me to do is to perform some statistical analyses on how the frequencies of *must* (and the other modals and semi-modals in my analysis) changed over time.

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.

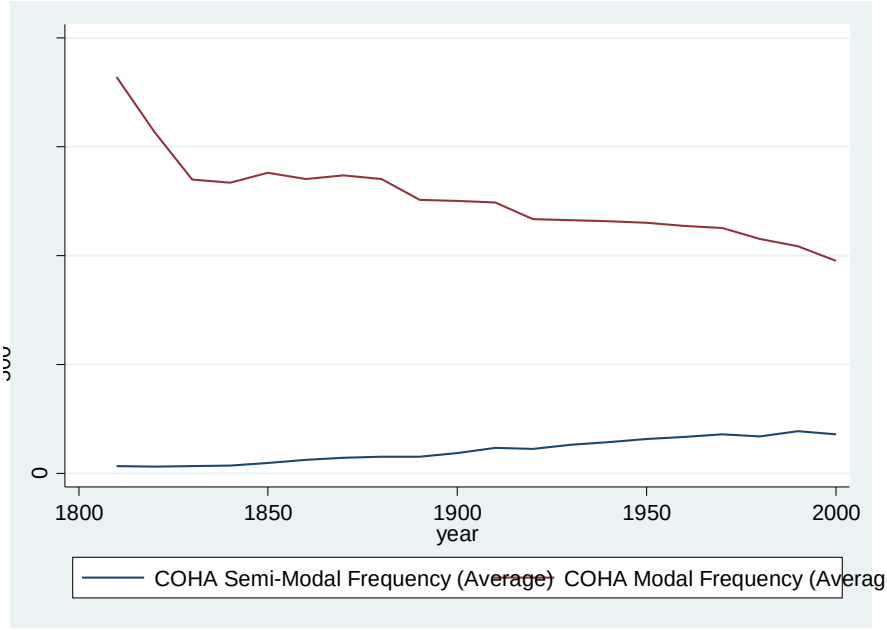






Note that, for the variable names in Stata and eViews, THC refers to the Hansard Corpus, while COHA is the Corpus of Historical American English. The names of the variables can be more easily understood once the prefixes are udnerstood, but there are some special terms that require explanation. Note 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.

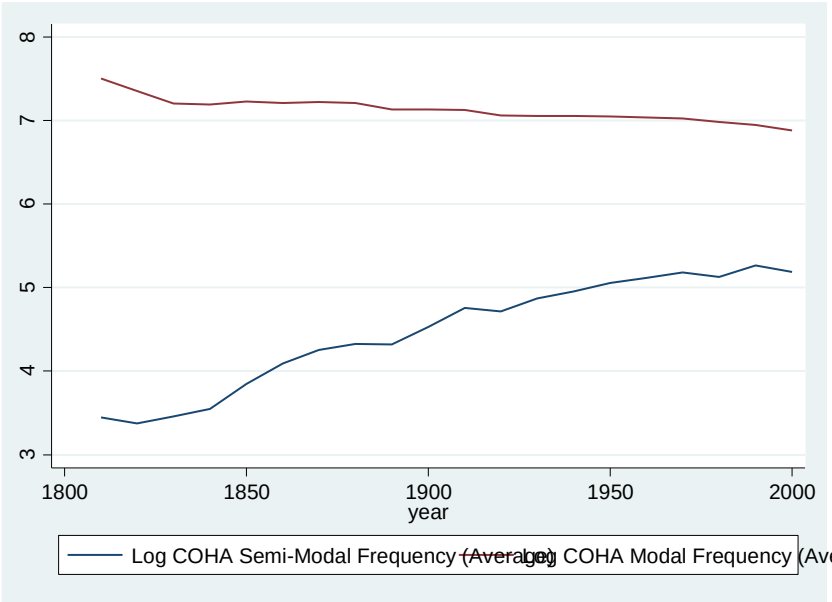
Note also that the ln prefix refers to a logarithmic transformation. Because raw frequencies were quite different (in terms of magnitude) between modals and semi-modals, I needed a method to compare them on the same graph. Using a log-transform procedure allowed me to achieve this kind of comparative 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 possoble 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:

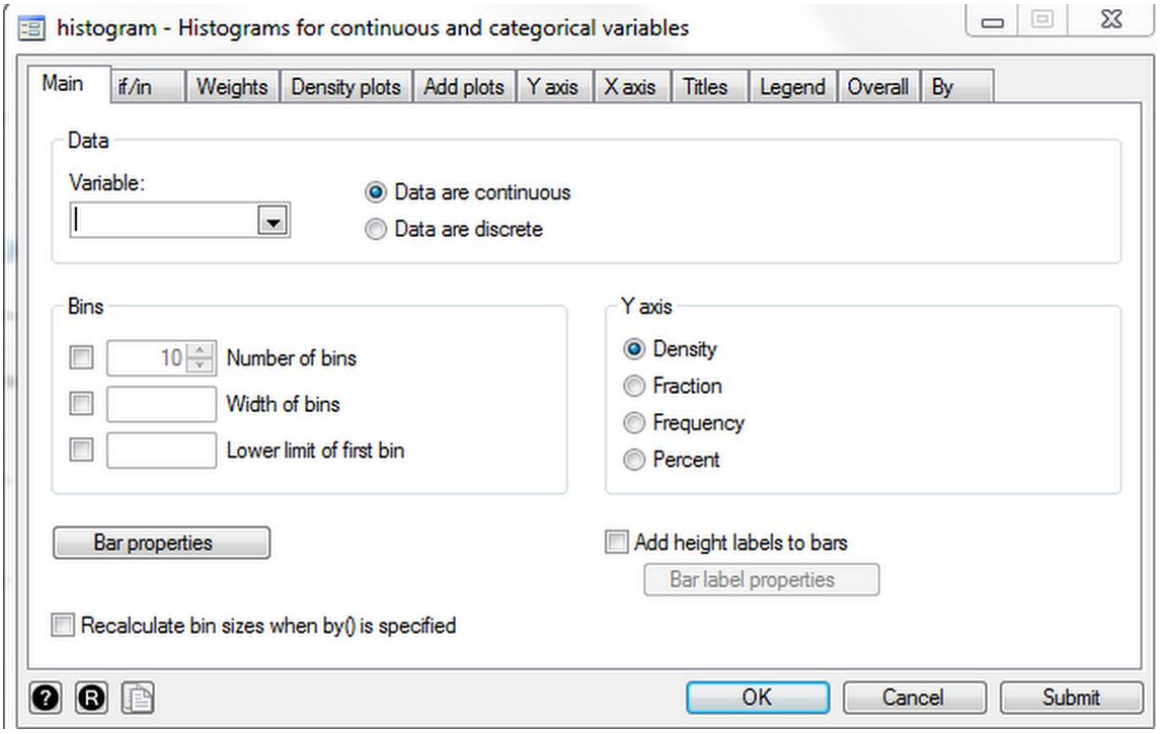
```
. reg coha_modal_avg coha_semimodal_avg
```

Source	SS	df	MS	Number of obs	=	20
				F(1, 18)	=	45.91
Model	504309.655	1	504309.655	Prob > F	=	0.0000
Residual	197736.128	18	10985.3404	R-squared	=	0.7183
				Adj R-squared	=	0.7027
Total	702045.783	19	36949.778	Root MSE	=	104.81

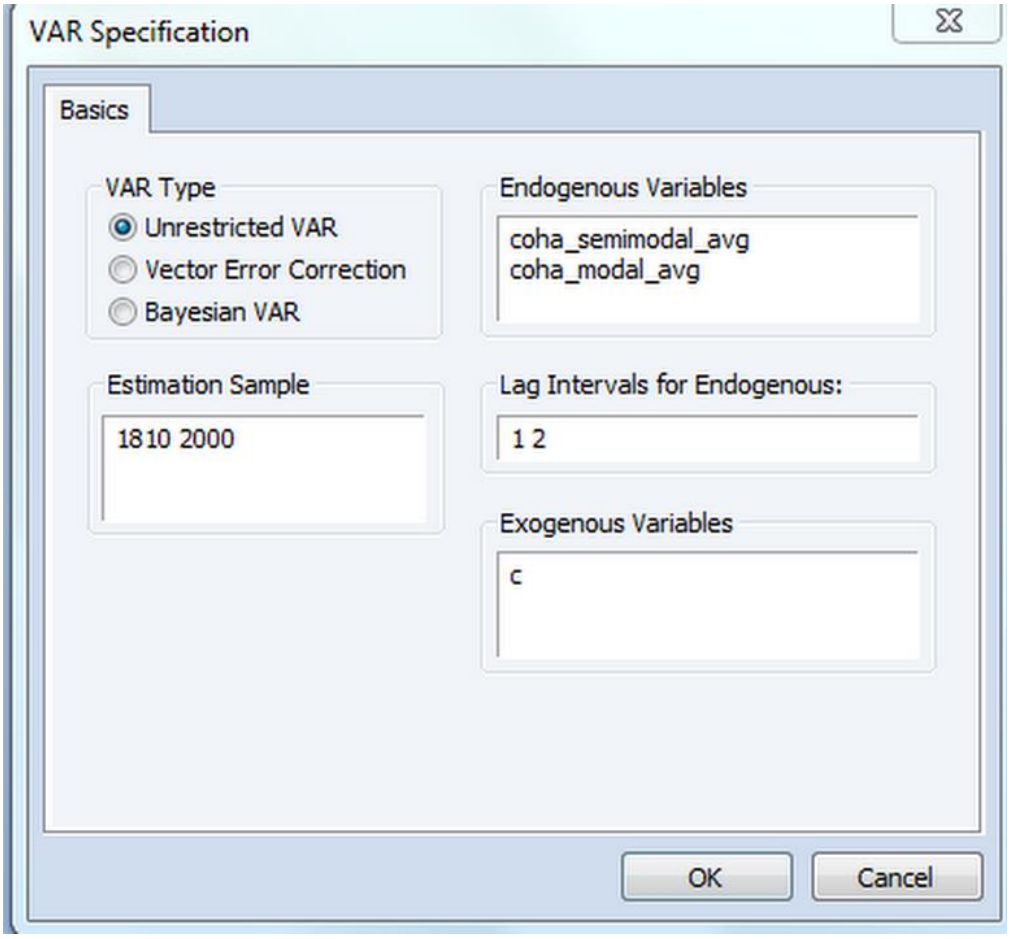
coha_modal_avg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
coha_semimodal_avg	-2.868013	.4232912	-6.78	0.000	-3.757315	-1.978711
_cons	1563.754	50.0822	31.22	0.000	1458.535	1668.973

These results were actually part of my study; a Ctrl-F search for 45.91 will identify where they occur in the document.

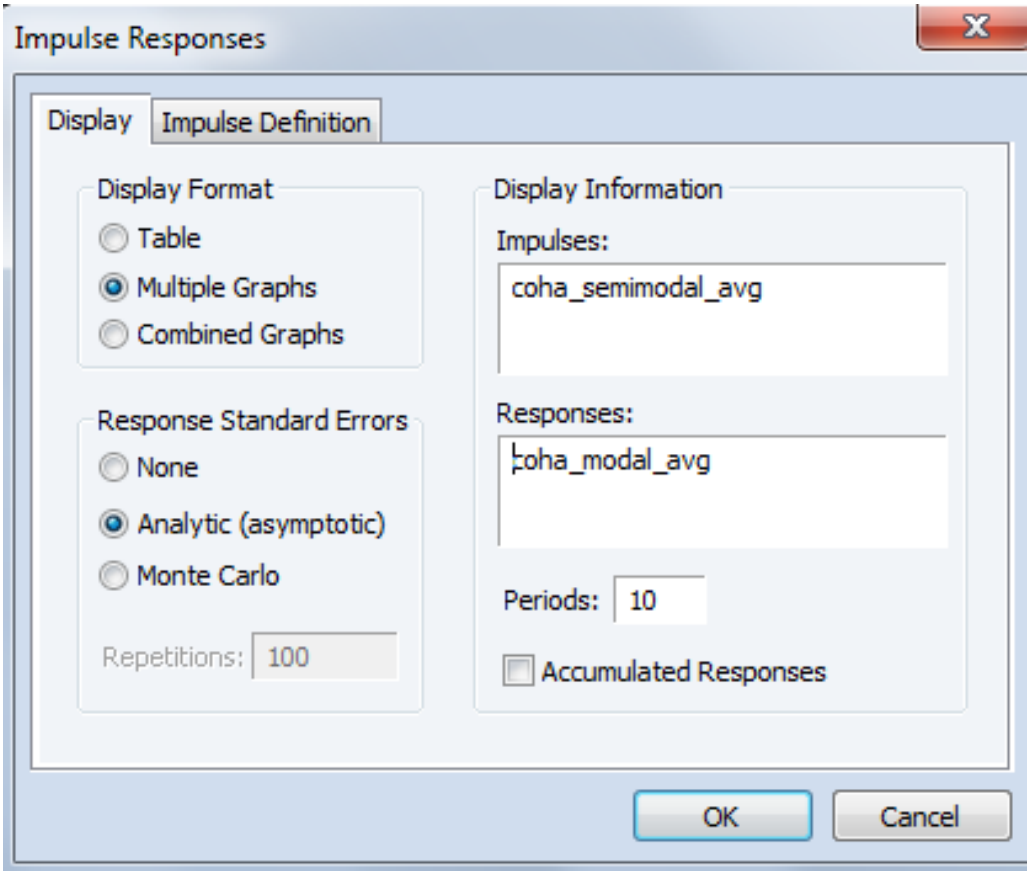
I created several histograms in Stata for inclusion in the thesis. The dialog box below demonstrates how histograms are created in Stata.



I used eViews for modeling the effect (as an impulse repsonse function, or IRF) of large changes in semi-modals on the subsequent frequencies of modals. I perfomed this analysis for hoth 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 autogression, 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.

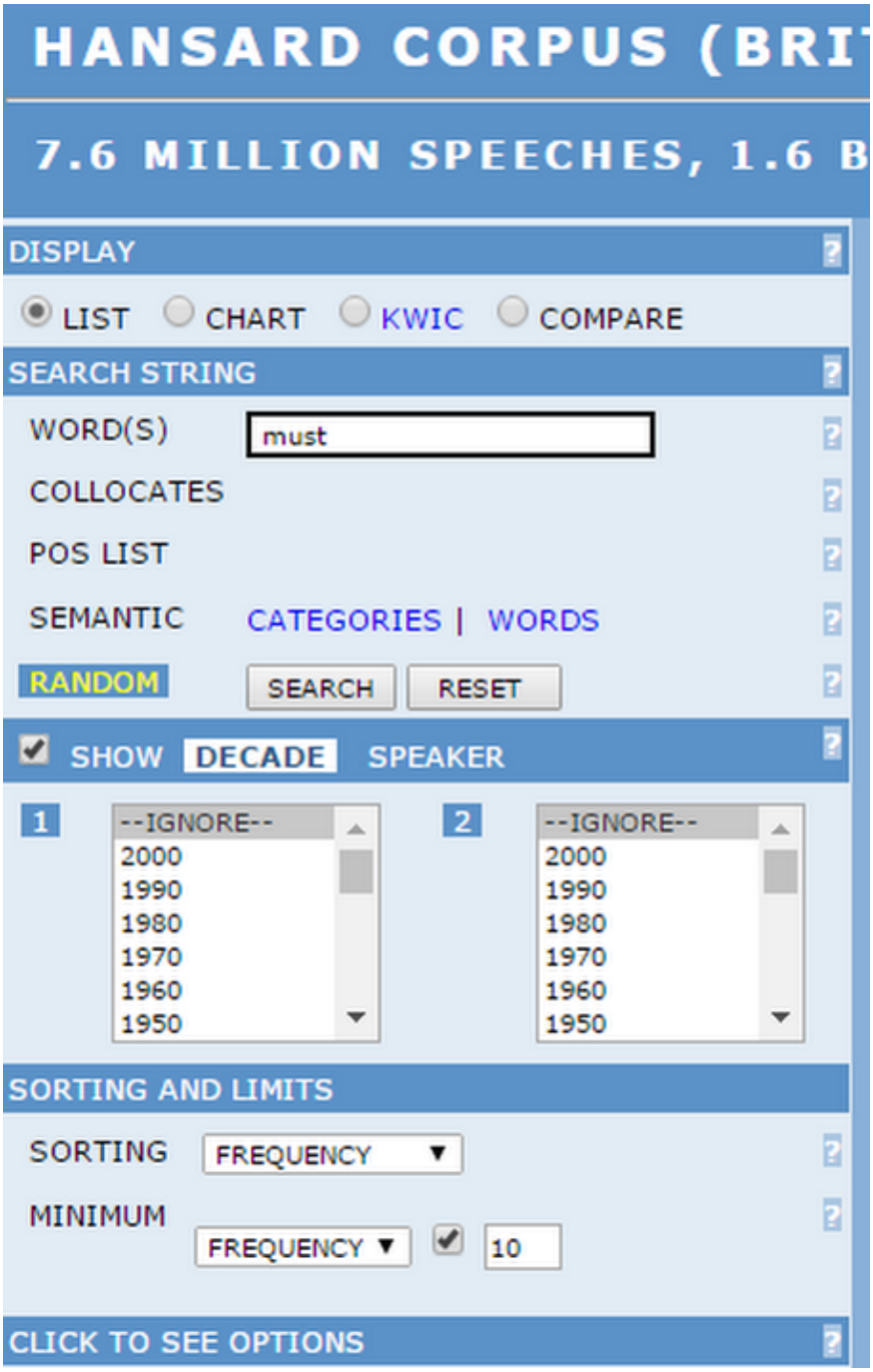
The same procedures that I used for COHA were used for THC as well. First, let me show how the Hansard Corpus can be retrieved:

Created by Mark Davies, BYU. **Overview**, search types, looking at variation, corpus-based resources, **upcoming**.

The most widely used online corpora -- more than 130,000 distinct researchers, teachers, and students each mo

English	# words	language/dialect	time period
Wikipedia Corpus (with virtual corpora)	1.9 **billion**	English	-2014
Global Web-Based English (GloWbE)	1.9 **billion**	20 countries	2012-13
Corpus of Contemporary American English (COCA)	**520** million	American	1990-**2015**
Corpus of Historical American English (COHA)	400 million	American	1810-2009
TIME Magazine Corpus	100 million	American	1923-2006
Corpus of American Soap Operas	100 million	American	2001-2012
British National Corpus (BYU-BNC)*	100 million	British	1980s-1993
Strathy Corpus (Canada)	50 million	Canadian	1970s-2000s
Hansard Corpus (British Parliament)	1.6 **billion**	British	1803-2005

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:

	year	thc_would	thc_will	thc_can	thc_could	thc_may
1	1810	7035.75	803.01	664.95	3683.55	612.1
2	1820	7626.87	857.87	665.94	3591.37	601.19
3	1830	8576.56	956.85	505.13	3181.62	502.96
4	1840	7740.32	1408.42	801.02	2835.05	761.19
5	1850	7536.76	1583.71	861.57	2533.69	900.64
6	1860	7481.59	1904.69	980.07	2314.56	1026.33
7	1870	7978.34	1693.96	854.22	2529.6	910.98
8	1880	7115.66	2654.35	1179.25	2068.7	1073.44
9	1890	5638.17	3954.82	1778.64	1486.36	1435.66
10	1900	5926.54	3688.26	1776.79	1749.73	1261.88
11	1910	3762.8	5544.18	2869.16	903.37	2056.1
12	1920	3577.49	5322.01	3055.82	926.35	2001.93
13	1930	3825.44	5589.6	3095.88	940	2045.79
14	1940	3854.65	6043.58	3324.96	1055.09	2047.28
15	1950	3881.74	5598.87	3243.51	1134.3	1919.57
16	1960	4230.43	5928	3146.74	1230.96	1851.33
17	1970	3971.55	6303.06	2991.04	1217.75	1758.5
18	1980	3714.47	6617.56	2886.8	1126.66	1567.11
19	1990	3574.61	6985.66	2832.59	1079.62	1558.84
20	2000	3709.58	6876.96	2881.58	1141.02	1509.06

Note that all of the raw data for the study have been presented in full in the body of the study itself.