RedFame

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

A Corpus Analysis of Changes in the Use of British and American English Modals and Semi-Modals

Abbas Hussein Tarish[1]

[1] Faculty of Foreign Languages and Literatures, University of Bucharest, Romania

Correspondence: Abbas Hussein Tarish, Faculty of Foreign Languages and Literatures, University of Bucharest, Romania.

Received: January 13, 2018 Accepted: February 5, 2018 Online Published: February 26, 2018

doi:10.11114/ijecs.v1i1. 3049 URL: https://doi.org/10.11114/ijecs.v1i1.3049

Abstract

This research has two main purposes. The first one is to test the *modal replacement hypothesis* proposed by Smith (2003) and discussed by Leech (2003), on the basis of data from the Hansard Corpus (THC- 1.6 billion words, 1800-2000) and the Corpus of Historical American English (COHA - 400 million words, 1810-2000). The second purpose of the study was to draw upon time series models to generate insights about how modal and semi-modal frequencies have changed over time. Cumulatively, these two forms of analysis addressed an acknowledged gap in the current literature on modal and semi-modal frequency change, namely the question of whether modals are being replaced by semi-modals.

Keywords: corpus analysis, modals, semi-modals, linguistic changes

1. Introduction

The word *modal* has several unique definitions in the Oxford English Dictionary. The definition of *modal* as applicable to grammar is as follows: "Of or relating to the mood of a verb; of a verb or other element; expressing or used to express modality. Especially in *modal auxiliary, modal verb*" (OED, 2016) The earliest usage of *modal* in its grammatical capacity cited in the Oxford English Dictionary was by J.H. Tooke, occurring in 1805. Penston (2012, p. 65) stated that a modal's purpose is "conveying the *mood* or opinion of the speaker, e.g. expressing ability, obligation, advice, possibility, etc.".

The word *semi-modal* does not appear in the Oxford English Dictionary. Dollinger (2006) defines semi-modals as verbs that share the characteristics of modal verbs and other verbs. Modals are ancient, with an analysis of the Oxford English Dictionary's etymological notes suggesting that the common modals of English can be traced back, through languages such as Saxon and German, all the way to Old Aryan. The antiquity of modal verbs can be understood as a natural feature of language, as the necessity of conveying mood or the opinion of the speaker is a basic function of language (Smith, 2003). The evolution of semi-modals, on the other hand, is more recent.

Modal verbs are:

- Would
- Will
- Can
- Could
- May
- Should
- Must
- Might
- Shall
- Ought
- Need

Semi-modal verbs are:

- Be going to
- Be to
- Have got to
- Have to
- Want to
- Used to

2. State of the Art

I will review the relevent literature into two parts. The first part consists of a focused discussion of the research by Leech (2003) and by Smith (2003), the linguists whose works are seminal, as well as being the closest in conception and content to this research. The second part consists of a general overview of the studies on modals and semi-modals frequency in English.

2.1 Seminal Work: Leech and Smith

One of Leech's (2003) main findings—based on the analysis of six corpora—was that of a decline in English modal auxiliaries from 1960 to 1990. Leech quantified this decline as being 10%. This analysis was based on the use of four corpora (LOB Corpus, Brown Corpus, SEU-mini-sp and ICE-GB-mini-sp), to measure the incidence of modal auxiliaries from around 1961, and two corpora (F-LOB Corpus and Frown Corpus) to measure the incidence of modal auxiliaries from around 1961. Leech described the space between the 1960 and 1990 corpora as constituting a generation gap.

Leech's (2003) analysis did not indicate strong support for the thesis, advanced primarily by grammaticalization theorists, that semi-modals have been displacing true modals. Leech's statistical analyses of British and American English indicated weak—at best—support for the modal-semi-modal competition theory. The empirical analysis of modal-semi-modal competition theory is complicated somewhat by the fact that there is controversy over the status of need to, which, according to Leech, is "arguably gaining [the] status" (Leech, 2003, p. 230) of a semi-modal. He noted what he characterized as a "remarkable rise" in the frequency of need to, with the difference between the frequency of this semi-modal (if it can indeed be considered a semi-modal) rising from 55 in LOB to 198 in FLOB, representing a 249.1% difference. Leech observed that, in American English, the frequency of need to increased 123.2% from the Brown corpus to the Frown corpus.

The change in relative frequencies of *need to* was the largest observed in Leech's analysis of 8 semi-models in 4 written corpora. It should be noted, in passing, that Leech's analysis was strongest when it was grounded on the written corpora, because of the absence of equivalent spoken corpora for American English and the existence of what he acknowledged to be mini-corpora (Leech, 2003, p. 231), for spoken British English from 1960.

Table 1. Leech's Observed Frequencies of Some Semi-Modals in 4 Written Corpora

	British English				American English			
	LOB	FLOB	Log-Likelihood	Diff (%)	Brown	Frown	Log-Likelihood	Diff (%)
BE *going to, gonna*	254	246	0.2	-3.1	219	332	23.5	51.6
BE *to*	454	376	7.6	-17.2	349	209	35.3	-40.1
(*had) better*	50	37	2.0	-26.0	41	34	0.7	-17.1
(HAVE) *got to, gotta*	41	27	2.9	-34.1	45	52	0.5	15.6
HAVE *to*	757	825	2.7	9	636	643	0.1	1.1
NEED *to*	55	198	83	249.1	69	154	33.3	123.2

WANT *to, wanna*	357	423	5.4	18.5	323	552	60.9	70.9
used to	86	97	0.6	12.8	52	71	3.0	36.5

Lecch's (2003) key findings from the written corpora are presented in Table 1 above, listed by corpus and nation and with log-likelihoods and percentage changes provided. For ease of comparison, table 2 below contains the semi-modal increases noted by Leech in the LOB-FLOB analysis and also semi-modal decrease in the DCPSE spoken corpus.

A comparison of the increases in semi-models between the chosen written corpus (LOB-FLOB) and the chosen spoken corpus (DCPSE), given in Table 2 below, reveals some interesting differences. In both of these corpora, the use of *have to, need to*, and *want to* increased. Note that *be going to, be to, have got to*, and *used to* fared differently in the written and spoken corpora. This finding indicates that there is only a partial overlap across written and spoken British English in terms of the pace of semi-modal usage frequency change.

Table 2. Comparison of Semi-Model Increases in Written and Spoken Corpora

	Written Corpus				Spoken Corpus			
	LOB	FLOB	Log-Likelihood	Diff (%)	DCPSE 1960	DCPSE 1990	Log-Likelihood	Diff (%)
BE *going to, gonna*	254	246	0.2	-3.1	1345	1778	25.25	31.3
BE *to*	454	376	7.6	-17.2	56	36	2.28	27.27
(had) better	50	37	2.0	-26.0	61	52	0.26	-13.2
(HAVE) *got to, gotta*	41	27	2.9	-34.1	443	444	0.0	0.5
HAVE *to*	757	825	2.7	9	1188	1307	2.47	10
NEED *to*	55	198	83	249.1	10	275	39.64	265
WANT *to, wanna*	357	423	5.4	18.5	837	1171	24.38	39.6
used to	86	97	0.6	12.8	305	276	0.27	-10.1

Leech's (2003) analysis of modals was limited to *may*, *should*, and *must*, with LOB and FLOB as the written corpora and SEU-mini-sp and ICE-GB-mini-sp (International corpus of English (Great Britain) spoken English texts from conversation BBC discussions, sports commentaries, other commentaries, BBC news, broadcast talks) as the spoken corpora. These results (Leech, 2003, pp. 232-233) are presented in Table 3 below.

Table 3. Leech's Observed Frequencies of Modals in 2 Written and 2 Spoken Corpora

	Written Corpora				Spoken Corpora			
	LOB	FLOB	Log-Likelihood	Diff (%)	SEU-mini-sp	ICE-GB-mini-sp	Log-Likelihood	Diff (%)
May	438	363	7.03	-20.66%	86	38	199.19	-1314.47%
Should	1301	1147	9.69	-13.43%	100	84	167.45	-644.05%

Must	382	270	19.34	-41.48%	87	35	208.84	-1453.57%

Smith (2003) advanced the thesis that the decline of modals was related to the rise of semi-modals. This argument is directly addressed in the empirical findings of the current study, presented in Section 4. Leech (2003) suggested that the forces of Americanization, democratization, and colloquialization might be responsible for the replacement of modals by semi-modals. One of Smith's (2003) points about the possible relationship between the decline of modals and the rise of semi-modals was that "a growing trend in published written discourse towards informal and less hierarchical styles" (Smith, 2003, p. 241) was responsible. If this explanation is correct, then modal styles considered formal and hierarchical might have been replaced by their less formal and hierarchical semi-modal equivalents. However, as the analysis of the Hansard Corpus demonstrates in Section 4, no such replacement can be seen when the analysis is expanded to the decades between 1810 and 2000. Moreover, the kind of 2-point cross-sectional analysis carried out by Leech (2003) cannot on its own supply the answer to the question of modal replacement; indeed, cross-sectional analysis is also fairly powerless to illustrate the longer trends in both modal and semi-modal frequencies. These gaps in the existing literature will be closed by means of the statistical approaches outlined in Section 3.

2.2 General Review of the Literature

The work of Leech (2003) has been enormously influential, with several other scholars having taken up the task of tracking the change in modal frequencies over time. One such scholar was Millar (2009), who used the TIME Magazine corpus to analyse changes in modal frequencies from 1923-2006. Millar's work contains an important observation, one that motivates the methodology of the present research : "a diachronic comparison based on two data points may present an inaccurate picture of the overall trend" (Millar, 2009, p. 191). Indeed, the regressions and time-series models in the current study demonstrate the usefulness of being able to examine modal and semi-modal change over a long span of time, with several data points, rather than the diachronic approach taken by Leech. In this sense, the current study is methodologically closer to Millar's work than to Leech's work. Even though Leech and Smith collaborated on an empirical paper (Leech & Smith, 2009), at around the time of the publication of Millar's paper, they continued to use older, cross-sectional methods of analysis rather than the technique advocated by Millar. Despite the existence of more advanced and explanatorily powerful techniques such as those used by Millar, several scholars (Berkenfield, 2006; Biber, 2004; Dollinger, 2006; Nokkonen, 2006; Rossouw and van Rooy, 2012; Vis, Sanders, and Spooren, 2012) have continued to prefer the diachronic techniques used by Leech (2003), Smith (2003), and Leech and Smith (2009).

Millar (2009) found a general rise in modal growth. Newspapers and journals have been leaders in terms of creating accessible virtual corpora, but the rise of large virtual corpora such as THC and COHA means that, in some respects, one-source corpora such as that of TIME are perhaps obsolete in corpus analysis. Thus, while Millar's work is an important demonstration of statistical techniques that go beyond diachronic analysis, its underlying corpus is not necessarily of high value.

Occasionally, the choice of limited corpora and diachronic analysis has led to conclusions about the relations between American and British English that are not well supported. For example, using diachronic analysis and drawing upon newspaper-based corpora for American English, Hundt reached the conclusion Hundt (1997) that the decline of *must* in American usage had influenced the decline of *must* in English usage. But this conclusion is not supported by an analysis of THC and COHA frequencies, in which *must* is observed to decline in both British and American English from 1810 to around 1920, after which the British use of *must* rises in comparison to the American use of *must*. Hundt's use of diachronic analysis prevented an identification of the actual pattern in the data, which only becomes apparent in the kind of time-series graphs used by Millar (2009).

3. Methodology

Consider the following modals and semi-modals, with *need* omitted because of what Leech (2003) described as its transition into semi-modal status:

Table 4. Tabulation Modals and Semi-Modals

Modals	Semi-Modals
Would	Be going to
Will	Be to

Can	Have got to
Could	Have to
May	Want to
Should	Used to
Must	
Might	
Shall	
Ought	
Need	

In this chapter, the changes in relative frequency of these modals and semi-modals will be tracked and interpreted, using 2 historical written corpora:

(a) The Hansard Corpus (THC), which contains 1.6 billion words and encompasses 7.6 million speeches made in the British Parliament from 1800 to 2000;

(b) The Corpus of Historical American English (COHA), a 400-million word corpus encompassing written American English from 1810 to 2000.

First, consider the thesis of semi-modal replacement. If semi-modals are replacing modals over a long sweep of historical time, then one means of detecting such change is to calculate the relative frequency of all modals or semi-modals in some corpora, and calculate log-likelihoods of changing frequencies based on cross-sectional comparisons. Such procedures have been followed by Leech (2003) and other linguists.

However, there are richer possibilities for analysis, given that THC and COHA are not monolithic, cross-sectional corpora, but rather time-series corpora. Each corpus offers the opportunity to calculate the relative frequency of either modals or semi-modals on a year-by-year basis. When yearly data points can be gathered, it is possible to use methods other than 2-point cross-sectional frequency comparisons to determine if there has been an ongoing replacement of modals with semi-modals.

For example, an ordinary least squares (OLS) model (Eisenhauer, 2003) can be fit to the data. An OLS model is a form of regression proceeding under the assumption that a 1-unit change in the independent variable can be associated with a fixed *n*-unit change in the dependent variable. The dependent variable is a variable that depends on one or more other variables, while an independent variable is a variable in an equation that may have its value freely chosen without considering values of any other variable. Independent variables are graphed on the horizontal axis and the dependent variables are graphed on the vertical axis (Eisenhauer, 2003). OLS models are often a default model for regression, as deviations from the OLS line of best fit are easy to see on a scatterplot, and because OLS models' significance levels, coefficients of determination, and Beta coefficients are easy to obtain and interpret (Eisenhauer, 2003). Linear models can be fit to determine whether patterns of rise and decline are observable on a year-by-year basis.

Another example of more informative statistical techniques is an unobserved components model (UCM) (Morley, Nelson, & Zivot, 2003) that could be used to determine whether the frequency of any particular modal or semi-modal has grown or declined in a particular way (such as through a random walk with trend). UCM is a form of time series analysis utilized when (a) there are data that change over time and (b) there is no pre-existing explanation of what might be influencing changes in data. A UCM is described as *unobserved* because it is not clear what variables might be influencing changes over time (Morley et al., 2003).

Finally, Markov-switching techniques can be used to detect specific historical eras (perhaps spanning multiple decades) in which there are marked rises or falls in frequencies of either modals or semi-modals (Kim, 1994). The basic assumption of Markov-switching is that the same time-series can contain multiple eras, that is, periods during which the dynamics of change are different (Kim, 1994). For example, if semi-modals were observed to increase from 1800 to 1900, and then decline from 1900 to 2000, Markov-switching would identify 2 eras in the data.

Because there are numerous modals and semi-modals in the study, and because of space restrictions in the thesis, some of the time-series procedures have been carried out on aggregate frequencies. However, the methods demonstrated with aggregate frequencies (that is, the frequency of all semi-modals or all modals taken as a class) can be just as easily utilized on individual modals or semi-modals.

4. Results

Chapter 4 has been divided into five parts. The first part of chapter 4 consists of a presentation of raw data. The second part presents results obtained by using regression models for both corpora. The third part consists of time-series analysis. The fourth part addresses distributions and cross-sectional analysis.

4.1. Raw Data

The four tables below show the frequencies of modals and semi-modals in THC and COHA corpora, which I extracted from data taken every decade from 1810 to 2000. Thus, we obtained 20 data points spanning 190 years.

The modal frequencies from THC are presented in table 5 and the semi-modal frequencies from THC are presented Table 6, while the modal frequencies from *COHA* are presented in table 7 and the semi-modal frequencies from *COHA* are presented Table 8.

Table 5. Modal Frequencies, THC

year	Thc Would	thc will	thc can	thc could	thc may	thc should	thc must	thc might	thc shall	thc ought	thc need
1810	7035.75	803.01	664.95	3683.55	612.1	3540.02	1372.5	2498.87	347.75	897.06	60.13
1820	7626.87	857.87	665.94	3591.37	601.19	3305.84	1416.78	2353.52	339.86	954.48	60.27
1830	8576.56	956.85	505.13	3181.62	502.96	3828.2	1396.39	2211.12	386.52	925.5	75.47
1840	7740.32	1408.42	801.02	2835.05	761.19	3421.28	1383.4	2050.62	479.82	799.71	79.25
1850	7536.76	1583.71	861.57	2533.69	900.64	3685.57	1289.61	2046.64	503.63	871.51	89.14
1860	7481.59	1904.69	980.07	2314.56	1026.33	3591.05	1173.65	1941.05	497.57	848.32	103.6
1870	7978.34	1693.96	854.22	2529.6	910.98	3676.51	1145.34	2058.15	506.22	802.84	122.26
1880	7115.66	2654.35	1179.25	2068.7	1073.44	3204.22	980.95	1562.03	582.44	700.47	105.06
1890	5638.17	3954.82	1778.64	1486.36	1435.66	3043.91	934.85	1131.8	767.22	712.94	129.46
1900	5926.54	3688.26	1776.79	1749.73	1261.88	2792.42	903.79	1177.75	557.89	658.09	146.12
1910	3762.8	5544.18	2869.16	903.37	2056.1	2681.96	959.16	705.32	931.14	663.22	186.92
1920	3577.49	5322.01	3055.82	926.35	2001.93	2570.57	927.15	650.71	872.97	569.15	227.32
1930	3825.44	5589.6	3095.88	940	2045.79	2674.92	1044.36	698.48	892.81	534.04	279.22
1940	3854.65	6043.58	3324.96	1055.09	2047.28	2932.4	1255.1	687.36	960.58	418.55	379.55
1950	3881.74	5598.87	3243.51	1134.3	1919.57	3276.61	1247.47	710.29	883.74	394.13	441.35
1960	4230.43	5928	3146.74	1230.96	1851.33	3289.48	1364.59	736.01	839.11	299.2	569.62
1970	3971.55	6303.06	2991.04	1217.75	1758.5	3235.9	1351.06	657.29	1078.63	207.26	659.84
1980	3714.47	6617.56	2886.8	1126.66	1567.11	3099.59	1339.79	599.19	1155.67	136.5	743.57
1990	3574.61	6985.66	2832.59	1079.62	1558.84	2846.38	1274.21	551.03	1099.71	82	871.61
2000	3709.58	6876.96	2881.58	1141.02	1509.06	2629.55	1166.87	598.54	1006.65	71	1096.5

Table 6. Semi-Modal Frequencies, THC

Year	thc be going to	thc be to	thc have got to	thc have to	thc want to	thc used to
1810	0	107.23	0	58.73	3.5	16.82
1820	0.09	92.82	0.17	62.34	4.31	20.84

1830	0.07	105.43	0.14	82.73	9.34	19.56
1840	0.16	115.99	0.2	107.72	19.36	23.84
1850	0.24	131.06	0.39	155.92	30.25	22.52
1860	0.2	129.08	0.94	194.43	39.75	29.78
1870	0.27	128.87	0.59	239.13	43.29	32.94
1880	0.2	101.51	1.72	318.32	76.49	33.34
1890	0.45	106.57	5.84	405.86	122.28	31.37
1900	0.4	91.71	11.52	411.26	123.38	39.35
1910	1.01	74.87	35.61	536.37	370.4	47.31
1920	1.03	69.1	28.34	568.98	486.31	55.42
1930	1.08	73.41	15.05	647.89	460.18	66.28
1940	1.04	66.6	20.26	694.99	549.29	73.87
1950	1.49	70.19	15.74	665.22	557.15	76.66
1960	1.85	79.02	10.08	662.78	504.68	73.24
1970	1.95	78.47	5.78	676.95	456.68	73.95
1980	1.63	75.91	4.29	630.64	384.71	86.77
1990	1.08	69.52	1.59	603.6	433.88	99.58
2000	1.02	67.82	3.34	618.41	619.36	101.59

The modal frequencies from COHA are presented below:

Table 7. Modal Frequencies, COHA

year	coha_ would	coha_ will	coha_ can	coha_ could	coha_ may	coha_ should	coha must	coha might	coha shall	coha ought	coha need
1810	3025.72	4271.06	2509.3	1025.22	2389.09	1591.6	1860.81	845.75	1924.31	430.07	165.09
1820	3042.22	3069.36	2076.32	1637.46	1924.45	1649.88	1239.9	949.74	1259.82	229.1	154.46
1830	2538.08	2739.97	1715.55	1374.56	1698.05	1456.23	1139.27	829.86	1042.35	174.67	138.88
1840	2425.79	2724.32	1808.59	1407.18	1568.19	1357.21	1193.83	795.28	1067.46	167.49	167.31
1850	2703.68	2783.57	1787.86	1618.96	1548.05	1384.86	1197.63	844.66	933.97	184.13	207.2

1860	2706.38	2616.14	1780.78	1709.78	1408.34	1297.22	1209.68	832.84	907.54	188.16	225.51
1870	2959.55	2554.75	1792.78	1899.34	1262.29	1299.52	1224.53	809.81	790.58	232.19	240.38
1880	2939.02	2584.42	1798.39	1807.49	1288.44	1274.47	1178.63	758.32	788.35	212.64	245.28
1890	2736.2	2402.72	1618.14	1790.9	1152.28	1169.08	1077.38	720.89	720.16	179.41	240.67
1900	2737.49	2454.48	1646.51	1733.04	1219.86	1071.88	1128.04	733.29	608.94	179.02	258.58
1910	2860.8	2251.48	1716.65	1834.71	1123.98	1032.22	1164.64	736.68	476.15	204.4	294.71
1920	2768.76	2350.61	1578.21	1647.79	1041.6	934.13	966.06	722.97	404.7	156.66	260.27
1930	2976.88	2199.93	1522.64	1889.23	979.45	874.91	926.24	705.33	275.78	152.22	275.26
1940	3050.52	2118.86	1589.5	2000.66	931.95	806.06	901.68	686.71	234.15	129.54	294.19
1950	3060.69	2030.04	1717.67	2027.47	876.72	793.65	840.83	686.95	187.37	117.3	311.06
1960	3014.69	2010.53	1802.63	1978.75	891.39	765.73	802.89	668.88	152.23	89	324.77
1970	2968.57	1979.74	1839.62	1990.83	841.35	773.29	739.15	671.88	145.12	86.96	366.19
1980	2835.96	1864.91	1792.54	2081.85	789.9	668.51	619.17	631.3	109.97	65.93	386.87
1990	2603.9	1725.5	1929.78	2082.46	722.65	696.56	520.8	584.29	71.58	48.6	474.45
2000	2497.38	1559.66	1780.98	2053.75	604.52	647.5	399.53	594.81	53.27	37.07	513.47

Table 8. Semi-Modal Frequencies, COHA

Year	coha be going to	coha be to	coha have got to	coha have to	coha want to	coha used to
1810	0	62.65	2.54	58.41	30.48	34.71
1820	0.58	51.68	0.29	50.67	17.76	53.99
1830	0.15	46.24	1.6	53.87	29.84	58.73
1840	0.37	42.93	1.43	58.01	35.27	70.54
1850	0.49	43.59	1.82	74.01	65.32	96.17
1860	0.76	47.08	2.17	96.57	100.79	112.81
1870	0.86	42.99	3.77	122.67	129.73	121.54
1880	0.64	42.82	4.97	140.14	142.65	122.07
1890	0.97	38.78	3.59	152.47	138.78	117.42
1900	1.13	35.75	4.75	198.17	194.64	123.27

1910	1.45	38.06	5.81	257.97	260.12	135.5
1920	1.72	32	4.68	255.17	249.63	126.45
1930	1.87	30.77	4.11	318.83	274.85	151.77
1940	1.81	27.89	2.42	357.77	304.21	156.69
1950	1.67	30.64	2.85	406.32	343.17	158.36
1960	1.67	30.36	2.71	423.11	372.19	168.12
1970	1.97	26.75	2.9	451.77	406.34	176.15
1980	1.94	23.46	1.58	434.94	390.62	163.18
1990	1.22	22.19	1.57	478.21	453.63	203.75
2000	1.52	18.13	1.59	441.5	428.75	182.16

The data in the four tables above provided the basis for the analyses to follow.

4.2. Regression Models

The modal replacement thesis (Smith, 2003) is that, over time, semi-modals have come to replace modals. This claim can be empirically tested by regressing semi-modals frequency on modals frequency in an ordinary least squares (OLS) regression model:

THC modal frequency = βTHCsemi-modal frequency + β0

The first such model was fit for THC. Note that the OLS regression showed that there was a significant (at an Alpha of 0.10) effect of semi-modal frequency on modal frequency, $F(1, 18) = 4.22$, $p = 0.0549$. The equation was as follows:

THC Modal Frequency = (THC Semi-Modal Frequency)(0.429) + 1945.504

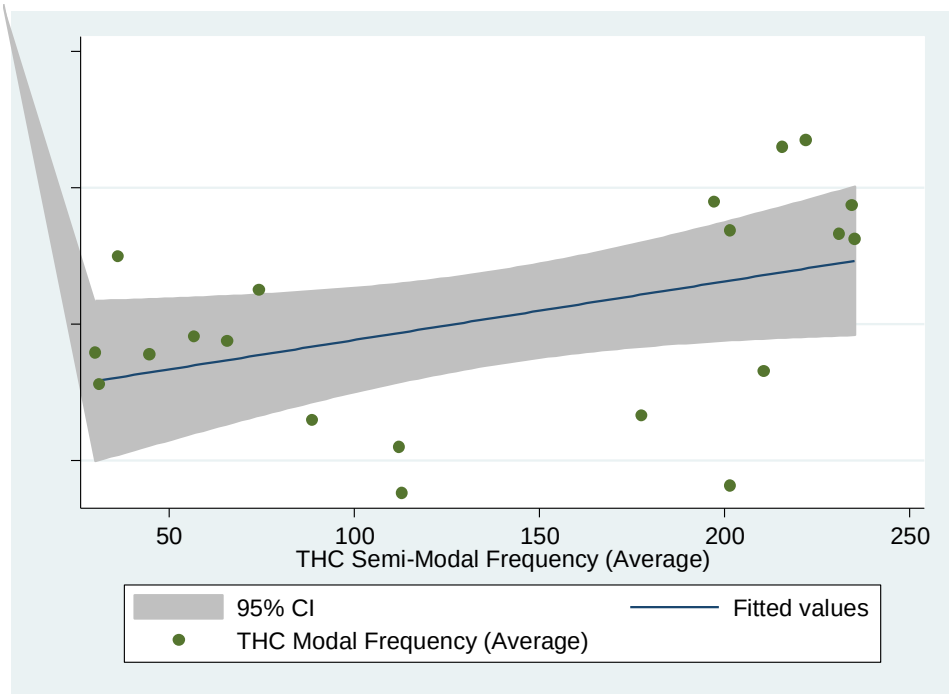

Figure 1. Scatterplot, relationship between modal and semi-modal frequency, THC. Note: OLS line of best fit and 95% confidence interval superimposed.

The coefficient of determination for the THC modals and semi-modals is 0.1897, which means that 18.97% of the variation in modal frequency can be explained by variation in semi-modal frequency. Note that the Beta coefficient for THC semi-modal frequency was positive, which is consistent with the interpretation that semi-modals are not replacing modals; rather, in THC, each 1-unit frequency increase in semi-modals is associated with a 0.429 increase in the frequency of modals. Therefore, it does not appear as if semi-modals are replacing modals in THC.

The same analysis was carried out with the COHA data. For COHA, as for THC, both modal and semi-modal frequencies were averaged and then placed into an OLS model:

COHA modal frequency = βCOHAsemi-modal frequency + β0

The scatterplot for this regression follows below. Note that the OLS regression showed that there was a significant (at an Alpha of 0.10) effect of semi-modal frequency on modal frequency, $F(1, 18) = 45.91$, $p < 0.0001$. The equation was as follows:

COHA Modal Frequency = (COHA Semi-Modal Frequency)(-2.868) + 1563.754

The coefficient of determination for the COHA modals and semi-modals is 0.7183, which means that 71.83% of the variation in COHA modal frequency can be explained by variation in semi-modal frequency. Note that the Beta coefficient for COHA semi-modal frequency was negative, which is consistent with the interpretation that semi-modals are replacing modals. In COHA, each 1-unit frequency increase in semi-modals is associated with a 2.868 decrease in the frequency of modals. Therefore, it appears as if semi-modals are replacing modals in COHA.

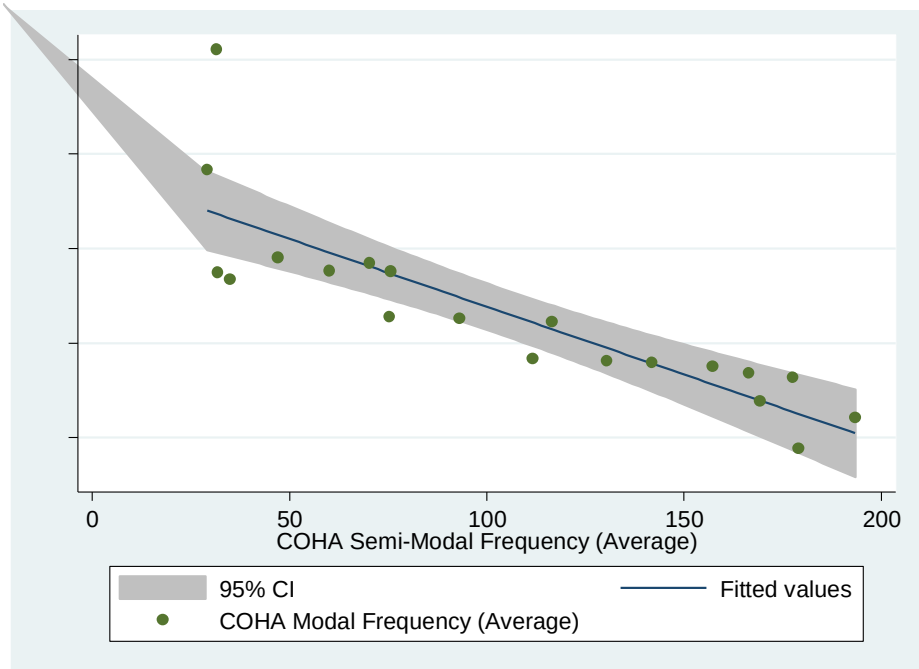

Figure 2. Scatterplot, relationship between modal and semi-modal frequency, COHA. Note: OLS line of best fit and 95% confidence interval superimposed.

One way of visualizing the possible replacement of modals by semi-modals in written American English is by placing modal and semi-modal frequencies on the same time line. This procedure does not work well with untransformed data, because modal frequency is far higher than semi-modal frequency and therefore does not provide useful information when placed on the same time line as semi-modal frequency. However, after both modal and semi-modal frequencies are log-transformed and placed on the same time line, they offer support for the claim that, in written American English, semi-modals are moving towards convergence with modals:

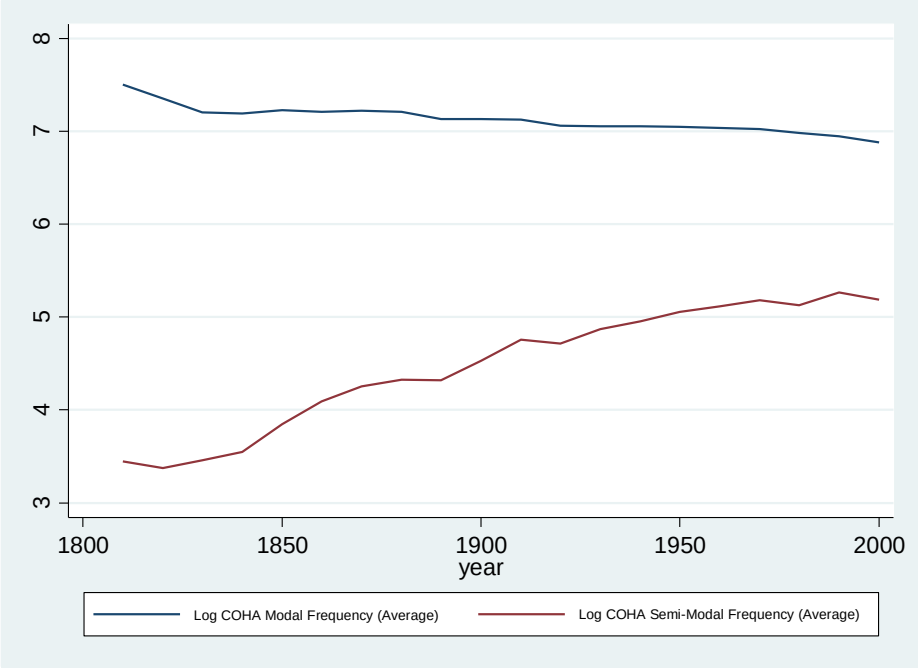

Figure 3. Change in log-transformed frequencies of modals and semi-modals in COHA, 1810-2000.

The results of the OLS models above are of special interest, as they indicate a possible effect of country on the modal replacement phenomenon. In American English, the COHA data from 1810 to 2000 support the replacement thesis; in British English, the THC data from 1810 to 2000 do not support the replacement thesis. This finding offers an empirical means to address the issue of modal replacement, which was left unresolved by both Leech (2003) and Smith (2003).

4.3. Time-Series Analysis

As mentioned in Section 3, the time-series nature of COHA and THC allow an investigation of the change in modal and semi-modal frequencies over time. The standard approach to the time-dependent comparison of corpora, as exemplified by Leech (2003), has been to compare a cross-section of a corpus at a particular point (such as 1960) to a cross-section of a corpus at another point (such as 1990). Cross-sectional comparison is a crude way of understanding change in a time series. A more robust and informative approach is to consider the change of a corpus in every year or every decade. In the case of THC and COHA, data were available for multiple years from 1810 to 2000. These corpora were sampled in every decade from 1810 to 2000 (a total of 19 data points), and the modal and semi-modal frequencies from these years were averaged.

The four time lines that follow illustrate the change in THC and COHA modal and semi-modal frequencies over time.

Although the time line for change in THC modal frequencies looks somewhat like a random walk, it can actually be fit with an OLS model that is significant, $F(1, 18) = 6.30$, $p = 0.0219$. The equation for this model is as follows:

THC modal frequency = (Year)(0.676) + 716.6431

The R^2 of this model is 0.2591, indicating that around 26% of the variation in THC's modals can be explained by the passage of time. Hence, every year, British English has added 0.676 in modal frequency. When a random walk model was fit on these data, Akaike's information criterion identified by (Allan and Chih 1998) "the Akaike information criterion, AIC, is measuring for regression and it is probably the most commonly used model selection criterion for time series data." Akaike's information criterion was actually lower for the random walk than for the linear model, suggesting that (a) there is a secular increase in THC modals and (b) the increase can be understood through the passage of time.

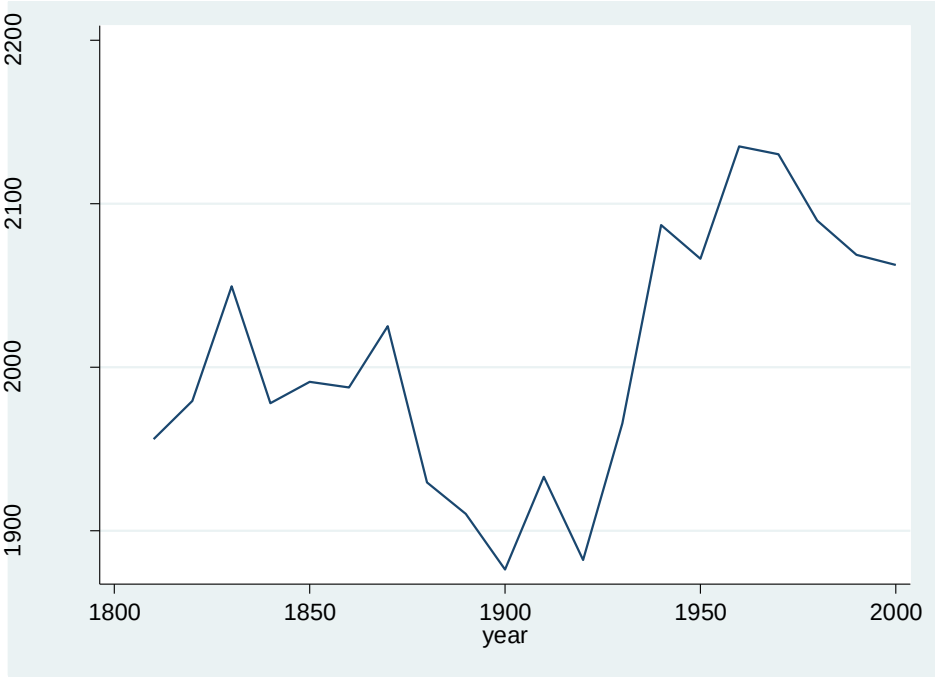

Figure 4. Time line, THC modal frequencies, 1810-2000.

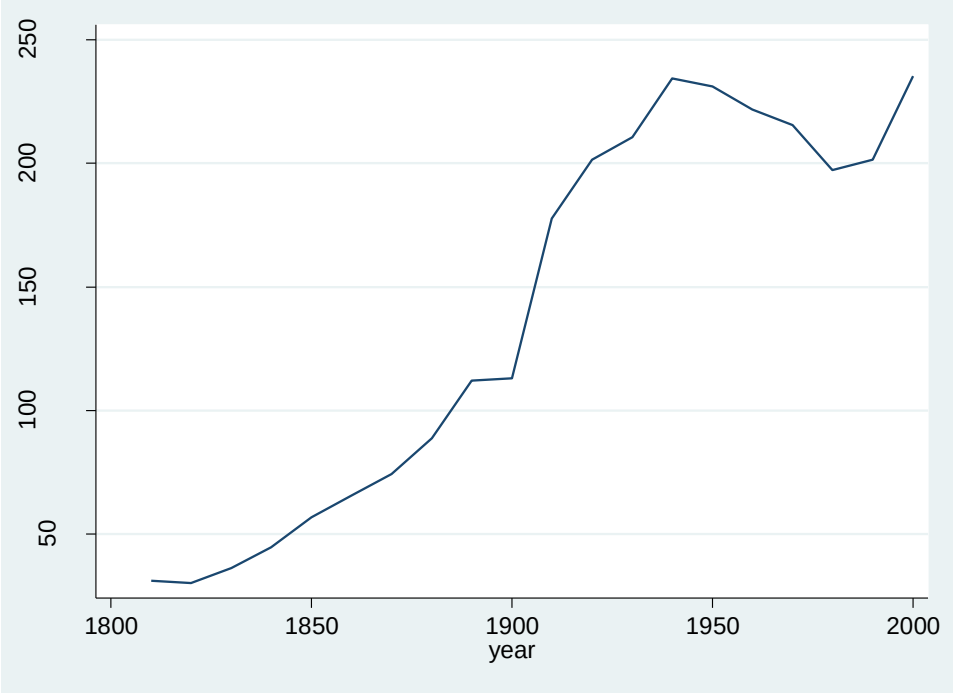

Figure 5. Time line, THC semi-modal frequencies, 1810-2000.

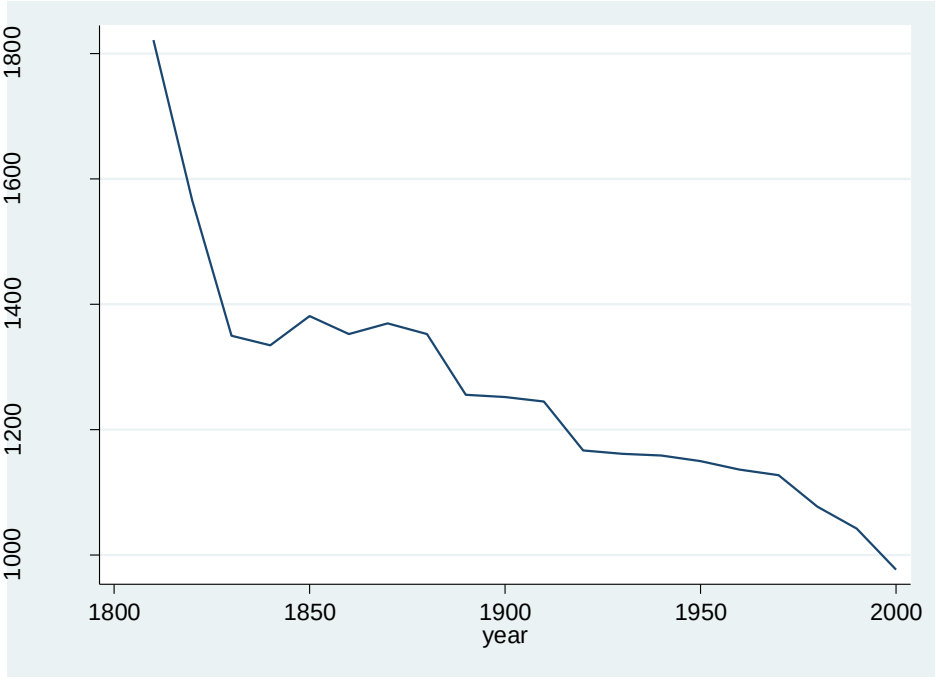

Figure 6 Time line, COHA modal frequencies, 1810-2000.

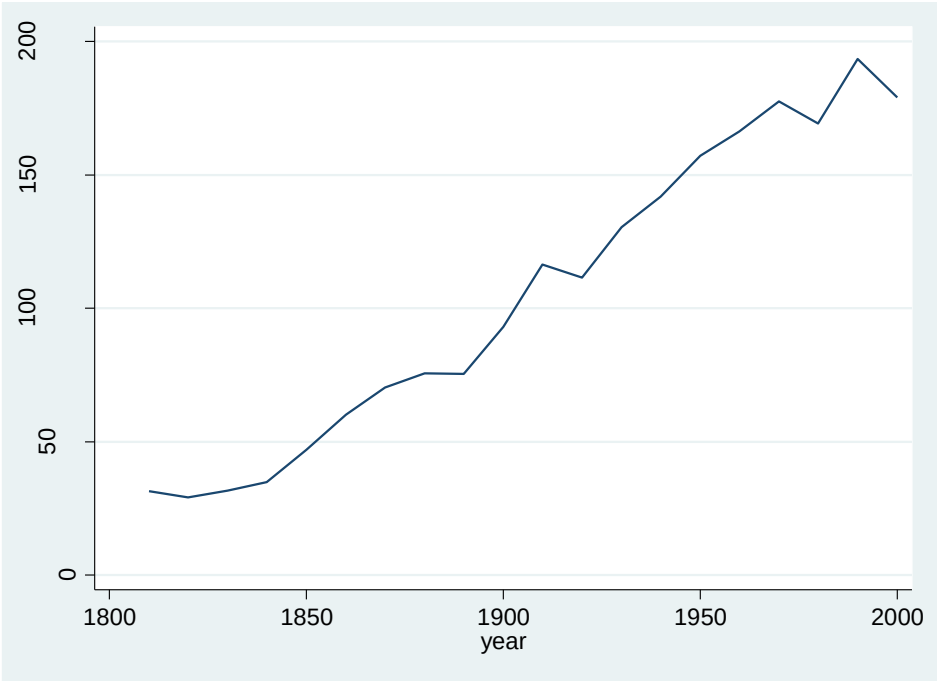

Figure 7. Time line, COHA semi-modal frequencies, 1810-2000.

The rise of semi-modals in THC was a good fit for an OLS model, $F(1, 18) = 131.39$, $p < 0.0001$, $R^2 = 0.8795$. The equation was as follows:

THC semi-modals = (Year)(1.263) – 2268.163

The decline of modals in COHA was a good fit for an OLS model, $F(1, 18) = 78.75$, $p < 0.0001$, $R^2 = 0.8140$. The equation was as follows:

COHA modals = (Year)(-2.931) + 6848.152

The rise of semi-modals in COHA was a good fit for an OLS model, $F(1, 18) = 706.53$, $p < 0.0001$, $R^2 = 0.9752$. The equation was as follows:

COHA semi-modals = (Year)(0.948) – 1701.734

Because OLS regressions provided excellent fits for 3 of the 4 time lines, and a significant but small fit for the 4th, no further UCM testing on Markov switching was attempted. The data appear to indicate that modals and semi-modals have both risen over time for THC, whereas, in COHA, modals have declined and semi-modals have increased.

The foregoing time series analyses offer insights into the nature of modal and semi-modal frequency change over time in both British and American English. One point of interest is to identify points of convergence and divergence between these two forms of English. The analysis of modals reveals that, over time, British and American usage has diverged significantly. American and British modals frequencies were nearly identical in 1810; however, by 1830, there was a fairly substantial gap in modal use between the 2 countries, a gap that has been becoming progressively wider. In terms of semi-modals, American and British written English again started out at about the same frequency. Then, starting at around the time of the First World War, the British use of semi-modals increased radically. At around the time of the Second World War, the British and American use of semi-modals once again began to converge, and, as of 1990, was nearly identical. Then, after 1990, another period of convergence seemed to have begun.

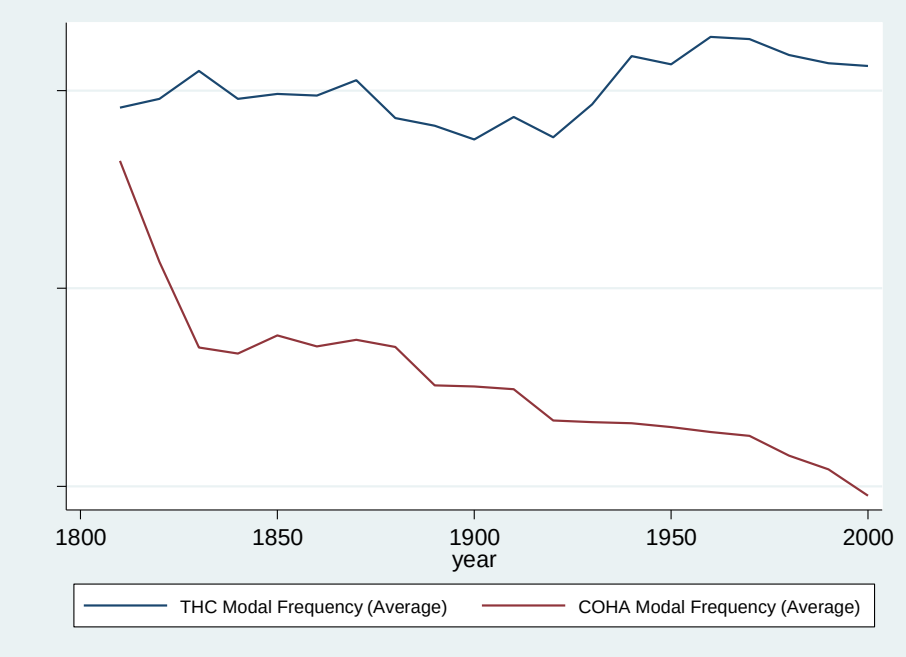

Figure 8. THC and COHA divergence, modal frequency.

Both Leech (2003) and Smith (2003) discussed the possible influence of American English on the increasing use of British semi-modals. However, a glance at Figure 8 confirms that, for nearly 2 centuries, written British English has higher semi-modal frequencies than written American English, a fact that does not appear to support the Americanization thesis.

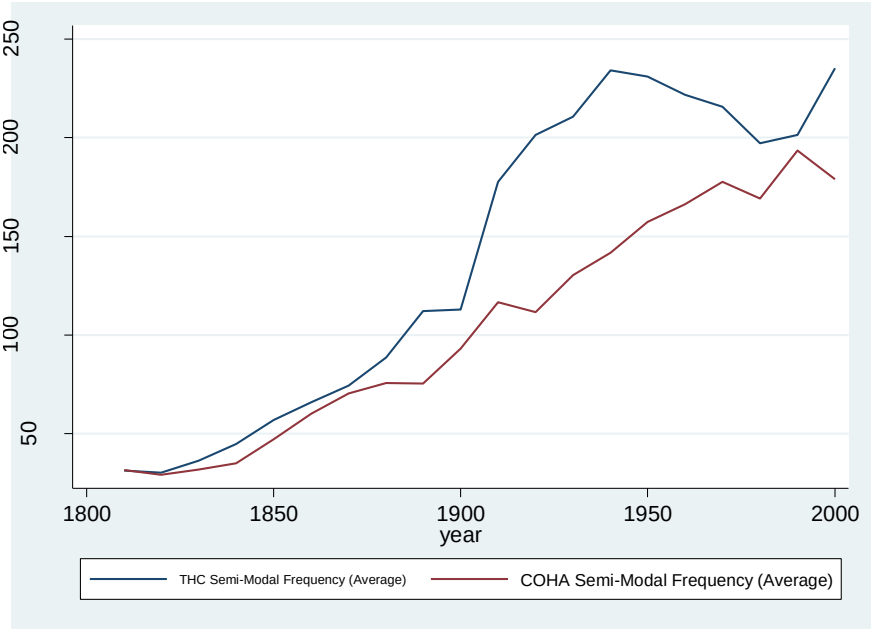

Figure 9. THC and COHA convergence, semi-modal frequency.

Of course, it should be noted that the analyses presented above are based on averaged figures. An examination of the frequencies of individual modals and semi-modals reveals more complex dynamics. For example, the fate of modal *would* demonstrates the opposite of the trend that can be seen in all modals (see Figure 9). With *would*, there is a convergence between British and American use, whereas, when all modals are averaged, there is a marked divergence between frequencies clearly visible in Figure 9.

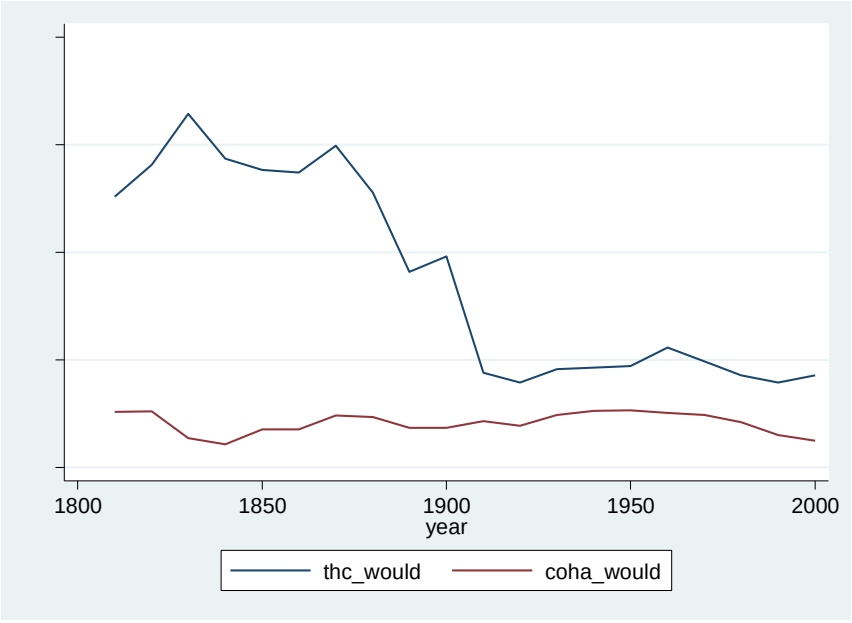

Figure 10. Frequencies for would in British and American written English over time.

In addition, simplistic ideas of divergence and convergence can be challenged by an analysis of the frequencies of modals such as *will*. *Will* was far more dominant in written American English at the beginning of the 19th century but was overtaken by the frequency of *will* in English usage at around 1870. One of the advantages of time-series analysis of frequency change, as opposed to the cross-sectional methods used by Leech (2003) and other scholars is that a full appreciation of the change in frequencies is possible. A 2-point cross-sectional analysis of the frequency of *will* could never reveal the full extent of how this modal has risen steadily in British English, while declining steadily in American English.

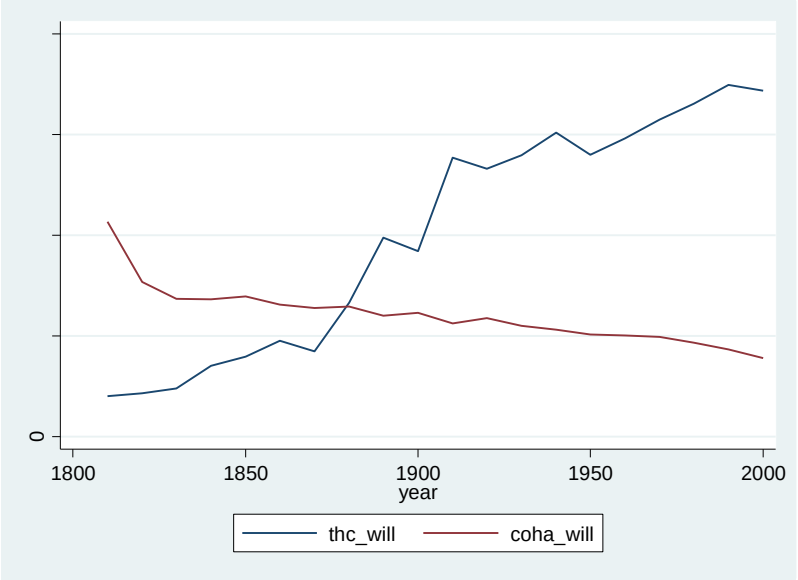

Figure 11. Frequencies for will in British and American written English over time.

When the modals and semi-modals are examined on their own, and in averaged terms, one of the key insights that emerge is that both modals and semi-modals simply became denser in British usage over the passage of time. The Americanization thesis is further challenged by this insight, as the data support the inference that, especially from the early years of the 20th century onwards, it is British English that has taken the lead in terms of denser usage of both modals and semi-modals. While the reason for this trend cannot be adduced from statistical analysis alone, its existence should serve as a guidepost to future scholars.

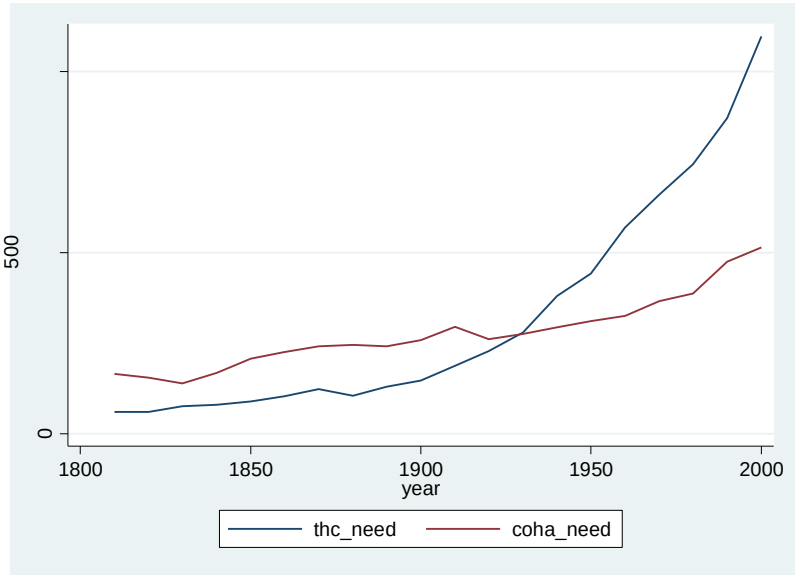

Figure 12. Frequencies for need in British and American written English over time.

Figure 12 illustrates that THC data *Need* indicated that the use of modals was substantially more frequent than COHA. *Need* was far more dominant in written American English at the beginning of the 19th century but was overtaken by the frequency of *Need* in English usage at around 1900.

The relationship between modals and semi-modals (addressing the modal replacement hypothesis) was explored in the regression models, but it is possible to extend this analysis to the impulse response function (IRF) technique in time series analysis. This technique (Box, Jenkins and Reinsel, 2011; Shumway and Stoffer, 2013) makes it possible to model the impact of a 'shock' in the independent variable to change in the dependent variable. For example, if a sudden, sharp rise in semi-modals occurred at one point in the recent history of American or British English, it would certainly be useful to be able to determine (a) just how long this shock influenced the frequency of semi-modals and (b) what the magnitude of the influence was.

The first IRF graph below is for COHA and suggests some interesting dynamics. The blue line is the element of interest; it represents the response of American modals to a change (specifically, an increase) in semi-modal frequencies. The red lines represent the 95% confidence interval of the blue line, which is best understood as the response of modals to the change in semi-modals.

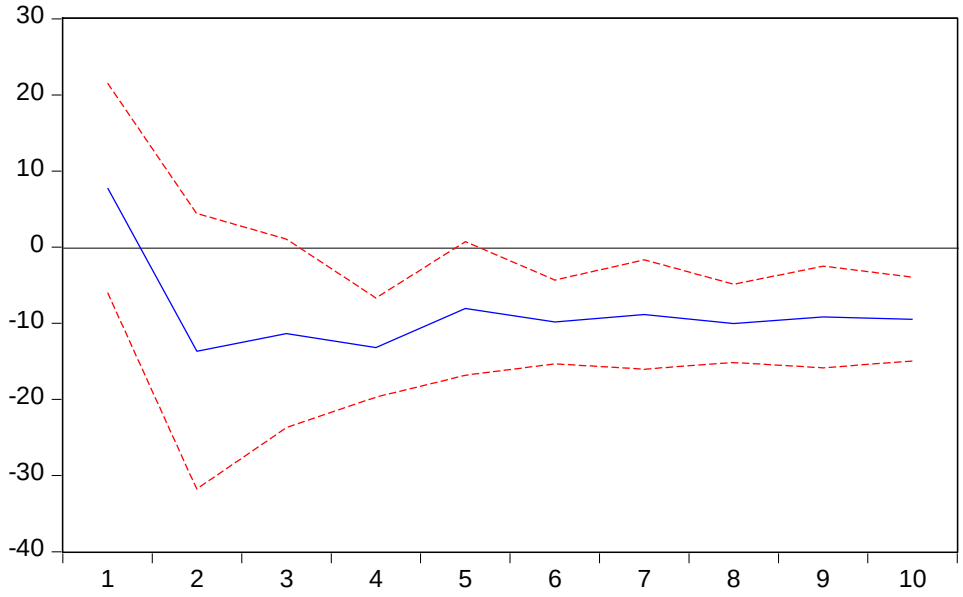

Figure 13. IRF Graph, impact of semi-modals on modals, COHA.

This graph offers a different way to think about the replacement thesis. Recall that, in Figure 1, it was showed that the increase in American semi-modals was negatively correlated with American modals. Also, figure 2 showed how a 1-standard deviation increase in semi-modals in American frequency takes only 2 decades to manifest itself as a decline in modal frequencies; note that 2 decades is the amount of time needed for the blue line in Figure 1 to go from positive to negative territory. Then, there is a century-long effect, as the blue line never reaches positive territory again. In other words, American English's response to an increase in semi-modals is to suppress modals 2 decades later, and the effect of this suppression lasts indefinitely.

The dynamics of the relationship between modals and semi-modals in British English are quite different, as the IRF illustration below conveys. In British English, an increase in semi-modals suppresses modal frequencies for about 30 years; however, after that, an increase in semi-modals actually leads to an increase in modal frequencies, and this effect lasts for 7 decades and more (note that both IRF graphs were programmed to stop after 10 decades' worth of analysis).

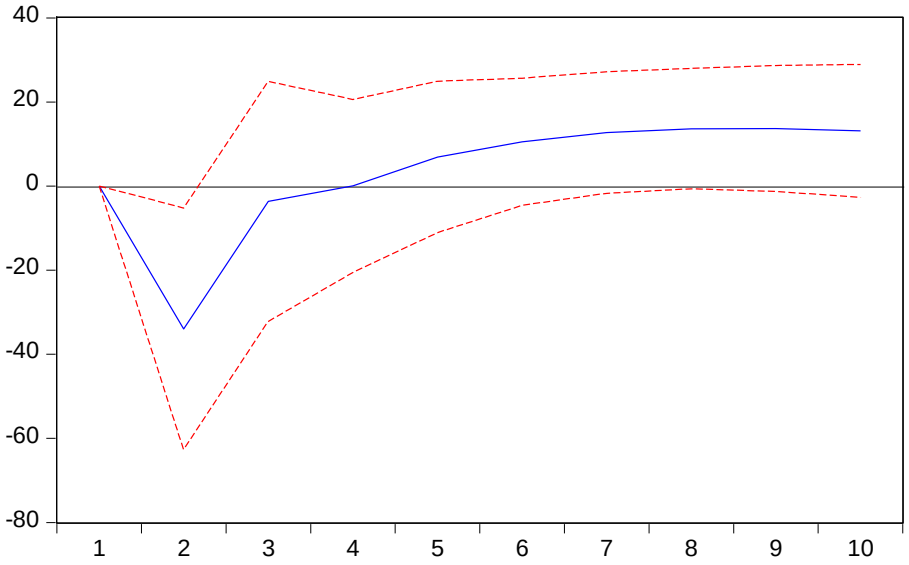

Figure 14. IRF Graph, impact of semi-modals on modals, THC.

The IRF graphs provide a form of insights that the regression models do not, but both the IRF graphs and the regression models illustrate the same basic phenomenon: the replacement of modals by semi-modals in English, and the joint increase in both modals and semi-modals in British English. The most important empirical contribution of the thesis is an identification of a replacement effect in one country but not in the other.

4.4. Distributions and Cross-Sectional Analyses

An analysis of histograms for modals and semi-modals in both THC and COHA offers yet another means of understanding the frequency distributions of modals and semi-modals in American and British English. An appropriate histogram with which to begin is the histogram of modal frequencies in COHA, which follows below:

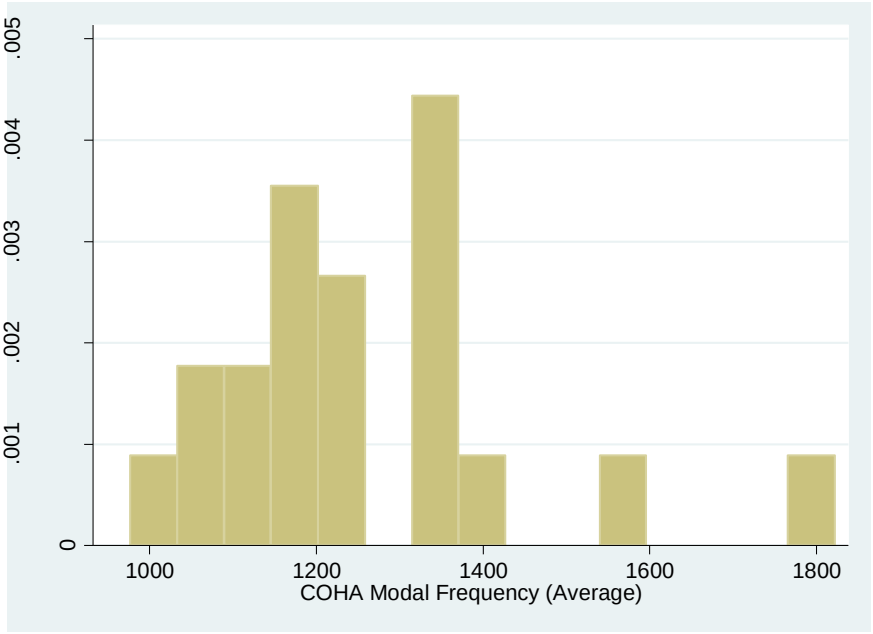

Figure 15. Histogram of modal frequencies, COHA.

Note that this distribution follows a Gaussian pattern. Modal frequencies peak between 1300 and 1400, and are less dense on the edges of the distribution. However, in the modal frequency distribution for THC, there is marked evidence of platykurtosis; the so-called flat distribution has a much tighter range (around 500 less than the range of the COHA

histogram) and supports the inference that British English has always had high usage of modals.

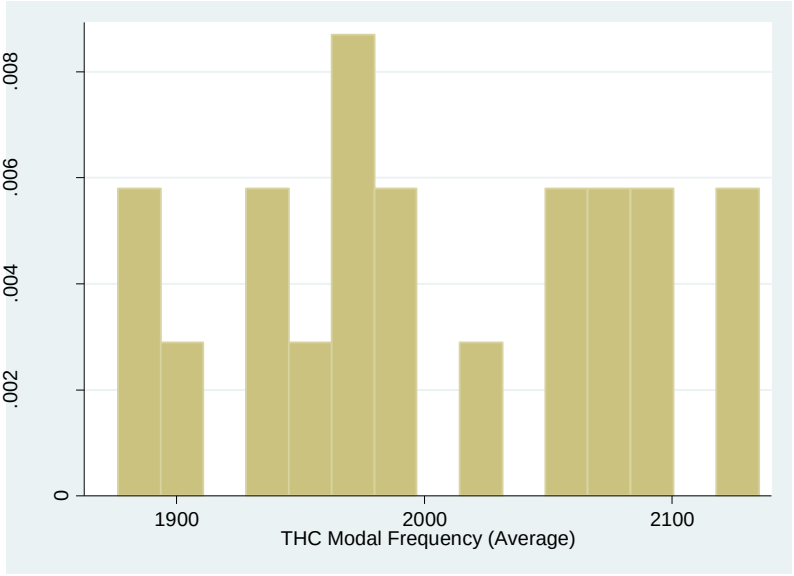

Figure 16. Histogram of modal frequencies, THC.

An analysis of the histograms for semi-modals for THC and COHA also illustrates a contrast between these 2 corpora. British English demonstrates peaks towards the right of the distribution, whereas, in American English, the highest peak is towards the left of the distribution. These differences in skewness offer some perspective on how semi-modal usage is simply denser in British English. The time-series analyses presented earlier confirm that the increasing density of semi-modal usage is dependent on time. However, one point for future scholars to address is why British English underwent, in the first half of the 20th century, a period of marked inflation of semi-modals as compared to American English.

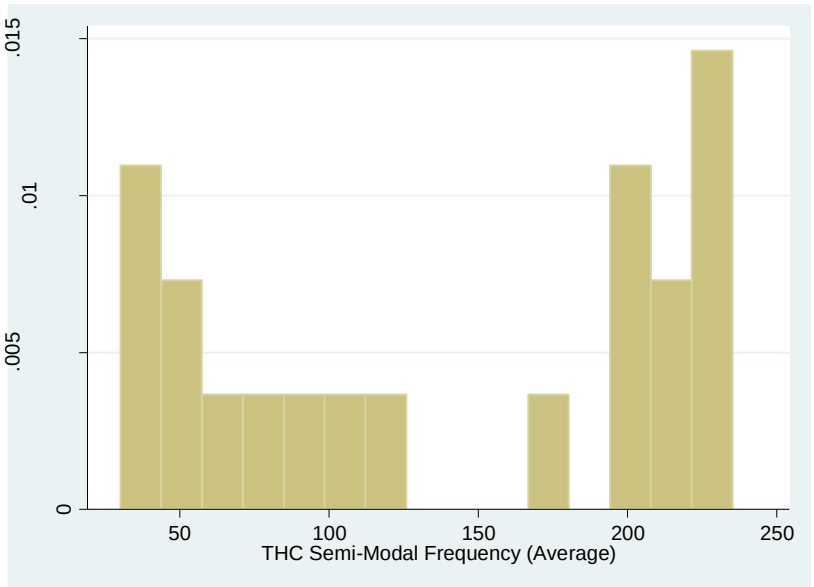

Figure 17. Histogram of semi-modal frequencies, THC.

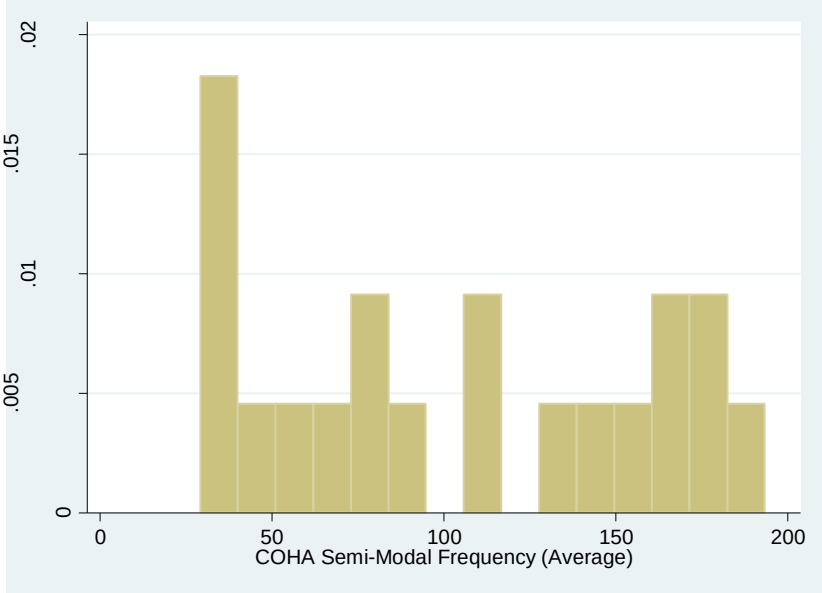

Figure 18. Histogram of semi-modal frequencies, COHA.

5. Conclusion

Leech's (2003) and Smith's (2003) work on semi-modals and modals was based on British English corpora and was delimited by time to the period between 1960 and 1990. Both Leech and Smith found evidence of a decline in modals and a rise in semi-modals. In the current study, this finding was confirmed only for American English, by means of COHA. An analysis of a historical British English corpus, THC, found that both modal and semi-modal frequencies rose over time. The results of the current study did confirm a rise in semi-modals for both American and British English, but a case for the existence of a possible substitution effect (in which the rise of semi-modals displaces the use of modals) can only be made with reference to American English. No such effect exists in THC.

In addition to providing analyses that addressed the issues of semi-modal increases and the modal substitution hypothesis, which have been discussed in past studies (Leech, 2003; Smith, 2003), this study also reached conclusions that are novel, in that they do not appear to have been discussed in previous empirical literature. One such conclusion pertained to modal divergence between American and British English. At the beginning of the 19th century, American and British English were fairly close together in terms of modal frequencies, but, over the course of the next 2 centuries, there was a steady rise in the modal frequencies of British English while modal frequencies in American English declined.

In terms of semi-modal frequencies, the time-series data suggest that, except for a relatively brief period during the first half of the 20th century, during which British usage of semi-modals outstripped American usage, there has been convergence. The convergence suggests that the Americanization hypothesis about the increase in semi-modals is incorrect, as least on the basis of THC and COHA. If anything, the data support a thesis of Anglicization, as British English had led American English in terms of semi-modal frequencies for nearly 200 years.

References

Berkenfield, C. (2006). Pragmatic motivations for the development of evidential and modal meaning in the construction "be supposed to X". *Journal of Historical Pragmatics, 7*(1), 39-71. https://doi.org/10.1075/jhp.7.1.03ber

Biber, D. (2004). Historical patterns for the grammatical marking of stance: A cross-register comparison. *Journal of Historical Pragmatics, 5*(1), 107-136. https://doi.org/10.1075/jhp.5.1.06bib

Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2011). *Time series analysis: forecasting and control*. New York, NY: John Wiley & Sons.

Charmaz, K. (2014). *Constructing grounded theory*. Thousand Oaks, CA: Sage.

Chih, L. T., & Allan, D. R. (1998). Regression and time series modal selection. World scintific, Co, Pte, Ltd. 4.

Dollinger, S. (2006). The modal auxiliaries have to and must in the Corpus of Early Ontario English: gradient change and colonial lag. *The Canadian Journal of Linguistics/La Revue Canadienne de Linguistique, 51*(2), 287-308.

Eisenhauer, J. G. (2003). Regression through the origin. *Teaching Statistics, 25*(3), 76-80. https://doi.org/10.1111/1467-9639.00136

Fischer, O. C. (2004). The development of the modals in English: radical versus gradual changes. In D. Hart (Ed.), *The development of the modals in english: Radical versus gradual changes*, 16-32. Bern, Switzerland: Peter Lang.

Hundt, M. (1997). Has BrE been catching up with AmE over the past thirty years? In M. Ljung (Ed.), *Corpus-based studies in English*, 134-150. New York, NY: Rodopi.

Kim, C. J. (1994). Dynamic linear models with Markov-switching. *Journal of Econometrics, 60*(1-2), 1-22. https://doi.org/10.1016/0304-4076(94)90036-1

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review, 73*(1), 31-43.

Leech, G. (2003). Modality on the move: The English modal auxiliaries 1961-1992. In M. G. King (Ed.), *Modality in contemporary English*. London, U.K.: Walter de Gruyter. https://doi.org/10.1515/9783110895339.223

Leech, G., & Smith, N. (2009). Change and constancy in linguistic change: How grammatical usage in written English evolved in the period 1931-1991. *Language and Computers, 69*(1), 173-200. https://doi.org/10.1163/9789042025981_011

Lightfoot, D. W. (1979). *Principles of diachronic syntax*. Cambridge, U.K.: CUP Archive.

Lorenz, D. (2012). *Contractions of English Semi-Modals: The Emancipating Effect of Frequency.* (PhD), Albert Ludwigs University, Freiburg, Freiburg.

Millar, N. (2009). Modal verbs in TIME: frequency changes 1923–2006. *International Journal of Corpus Linguistics, 14*(2), 191-220. https://doi.org/10.1075/ijcl.14.2.03mil

Morley, J. C., Nelson, C. R., & Zivot, E. (2003). Why are the Beveridge-Nelson and unobserved-components decompositions of GDP so different? *Review of Economics and Statistics, 85*(2), 235-243. https://doi.org/10.1162/003465303765299765

Nokkonen, S. (2006). The semantic variation of NEED TO in four recent British English corpora. *International Journal of Corpus Linguistics, 11*(1), 29-71. https://doi.org/10.1075/ijcl.11.1.03nok

OED. (2016). Modal. Retrieved from http://www.oed.com

Penston, T. (2012). *A concise grammar for English language teachers*. New York, NY: TP Publications.

Plank, F. (1984). The modals story retold. *Studies in Language, 8*(3), 305-364. https://doi.org/10.1075/sl.8.3.02pla

Roodman, D. (2007). The anarchy of numbers: aid, development, and cross-country empirics. *The World Bank Economic Review, 21*(2), 255-277. https://doi.org/10.1093/wber/lhm004

Rossouw, R., & van Rooy, B. (2012). Diachronic changes in modality in South African English. *English World-Wide, 33*(1), 1-26. https://doi.org/10.1075/eww.33.1.01ros

Seggewiss, F. (2012). *Current Changes in the English Modals--a Corpus-Based Analysis of Present-Day Spoken English.* (PhD), Albert Ludwigs University, Freiburg, Freiburg.

Shumway, R. H., & Stoffer, D. S. (2013). *Time series analysis and its applications*. New York, NY: Springer Science & Business Media.

Smith, G. (2003). Changes in the modals and semi-modals of strong obligation and epistemic necessity in recent British English. In M. G. King (Ed.), *Modality in contemporary English*. London, U.K.: Walter de Gruyter. https://doi.org/10.1515/9783110895339.241

Vis, K., Sanders, J., & Spooren, W. (2012). Diachronic changes in subjectivity and stance–A corpus linguistic study of Dutch news texts. *Discourse, Context & Media, 1*(2), 95-102. https://doi.org/10.1016/j.dcm.2012.09.003

Copyrights