Language Change

Seminar 'Corpus Linguistics'

Quirin Würschinger, LMU Munich

July 17, 2025

Session Overview

- Language change fundamentals
- Modal verbs in English: frequency changes over time Hilpert (2015)
- Practice: hands-on analysis using COHA and COCA on english-corpora.org

Language Change

What is language change?

- systematic modifications in language over time
- affects all linguistic levels: phonology, morphology, syntax, semantics

Examples

- **phonology**: *knight* /naɪt/ vs Middle English /kniçt/
- **morphology**: *you* (singular) replacing *thou/thee*
- **syntax**: *do*-support in questions (*Do you know?* vs *Know you?*)
- **semantics**: *nice* shifting from ‘foolish’ to ‘pleasant’

→ corpus linguistics provides empirical evidence for change

Research Questions in Language Change

- **What** changes? → linguistic features
  - → word order patterns: *do*-support in questions (*Do you know?* vs *Know you?*)
  - → use of modal verbs: *must* declining from 2000 to 500 per million words
- **When** does it change? → timing and pace
  - → *thou/thee* disappears rapidly in 17th century
  - → gradual decline of *shall* over 200 years
- **How** does it change? → mechanisms and patterns
  - → example: grammaticalization of *going to* → *gonna*
  - → concrete: *have to* replaces *must* in obligation contexts

- **Who** changes? → social factors
  - → factors: age groups, social classes, gender differences
  - → example: younger speakers use *gonna* more than older speakers
- **Why** does it change? → causes and motivations
  - → factors: social prestige, language contact, simplification
  - → example: *ain't* stigmatised, speakers switch to *isn't*

Corpus Methods for Language Change

- **Diachronic corpora**: e.g. EEBO, COHA, Gutenberg, COCA, NOW, English Trends
- **Frequency analysis**: absolute and relative frequencies over time
- **Text type variation**: register-specific changes
- **Collocation analysis**: changing semantic associations
- **Statistical measures**: coefficient of variation, significance testing

Modal Verbs in English

Theoretical Background: Hilpert (2015)

> Another domain of English grammar that is currently undergoing change is the domain of modality, specifically the modal auxiliaries. In the most general of terms, the situation is that several of the core modal auxiliaries are declining in text frequency (Leech 2003; Mair 2006), while at the same time new quasi-modal elements are undergoing grammaticalization (Krug 2000).

Key question: Why certain forms decline while others rise?

Core vs Peripheral Modal Verbs

Core Modal Verbs

- *will, would*
- *can, could*
- *may, might*
- *shall, should*
- *must*

Peripheral Modal Verbs

- *BE going to*
- *have to*
- *got to*
- *need to*

Frequency Changes Over Time

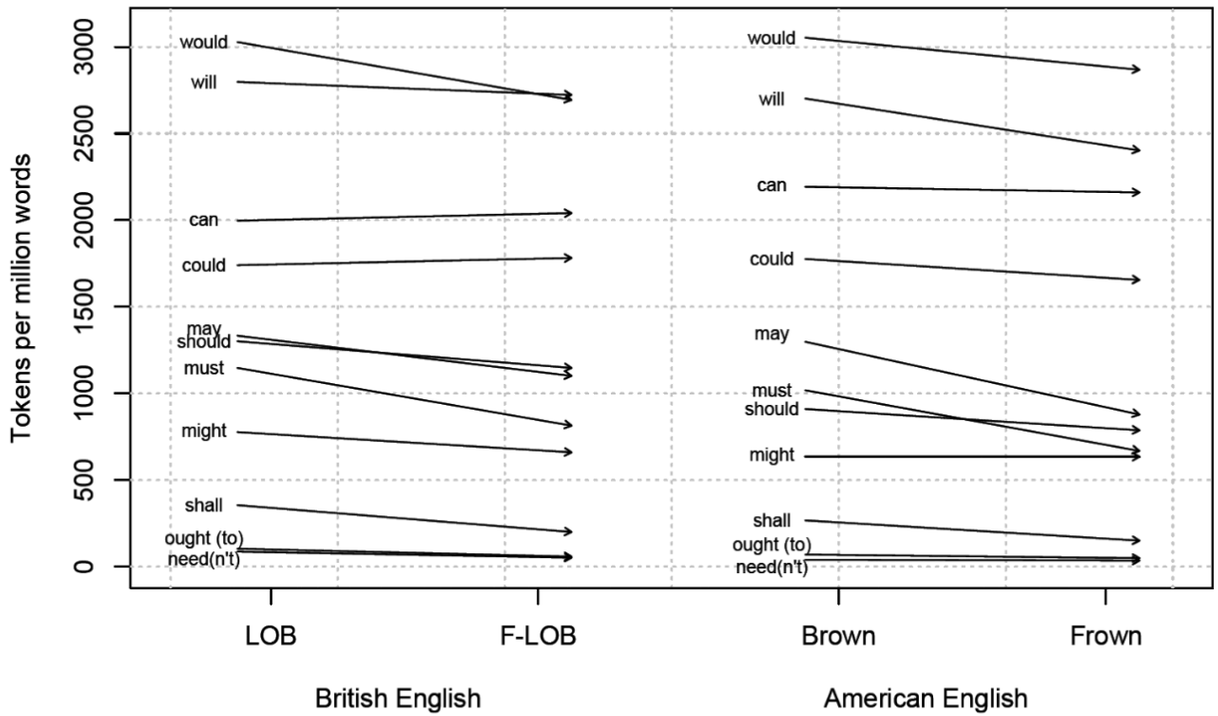


Figure 10.2 The decline of the modal auxiliaries (based on Leech 2003: 228, table 3)

Frequency changes of modal verbs over time.

Overall

- ↘️ **core** modals declining
- ↗️ **peripheral** modals rising

Interaction between frequency changes and causes

(Hilpert 2015, 186)

"The result is a dynamic situation that raises a number of questions.

- For instance, it has been asked **why** certain forms are in decline whereas others are on the upswing.
- Is there a relation between these developments, and if so, how do we assign the **roles of cause and effect**?"

Potential cause: text type variation

(Hilpert 2015, 187)

- “One explanation for the discrepancies between the tendencies in the Brown family of corpora and in the Time corpus is the **composition** of the respective corpora.
- Whereas the Brown corpora represent a **balanced** set of genres, the Time corpus represents a **single text type**.
- To test whether genre differences explain the discrepancies, Millar (2009: 207) compares his **Time** results against an analysis of the press genres in the **Brown** and **Frown** corpora, finding, however, no satisfactory convergence between the two.
- Millar thus invokes **sampling error** as an explanation, which is criticized by Leech (2011a), who replicates the results from the Brown family of corpora on the basis of the balanced **diachronic mega-corpora COCA and COHA** (Davies 2008, 2010).
- These results leave the frequency increases of *can*, *could*, and *may* in Time in need of an explanation, for which Leech (2011a: 557) suggests a **genre-specific style change** in journalistic writing.”

Practice: Corpus Analysis

Study objectives

1. **Frequency analysis**: track modal verb usage over time in COHA
2. **Text type variation**: examine register preferences in COCA

How to use english-corpora.org

Overview of Corpora

English-Corpora.org

corpora guides videos related resources users my account **upgrade** help

These are the most widely used online corpora, and they are used for many different purposes by teachers and researchers at universities throughout the world. In addition, the corpus data (e.g. full-text, word frequency) has been used by a wide range of companies in many different fields, especially technology and language learning.

The links below are for the free online interface. You can also purchase and download the corpora for use on your own computer.

Corpus	Overview	Download	# words	Dialect	Time period	Genre(s)
News on the Web (NOW)			17.5 **billion+**	20 countries	2010-yesterday	Web: News
iWeb: The Intelligent Web-based Corpus			14 **billion**	6 countries	2017	Web
Global Web-Based English (GloWbE)			1.9 **billion**	20 countries	2012-13	Web (incl blogs)
Wikipedia Corpus			1.9 **billion**	(Various)	2014	Wikipedia
Coronavirus Corpus			1.5 **billion**	20 countries	Jan 2020-Dec 2022	Web: News
Corpus of Contemporary American English (COCA)			1.0 **billion**	American	1990-2019	Balanced
Corpus of Historical American English (COHA)			475 million	American	1820-2019	Balanced
The TV Corpus			325 million	6 countries	1950-2018	TV shows
The Movie Corpus			200 million	6 countries	1930-2018	Movies
Corpus of American Soap Operas			100 million	American	2001-2012	TV shows

Views and Query Types

List View

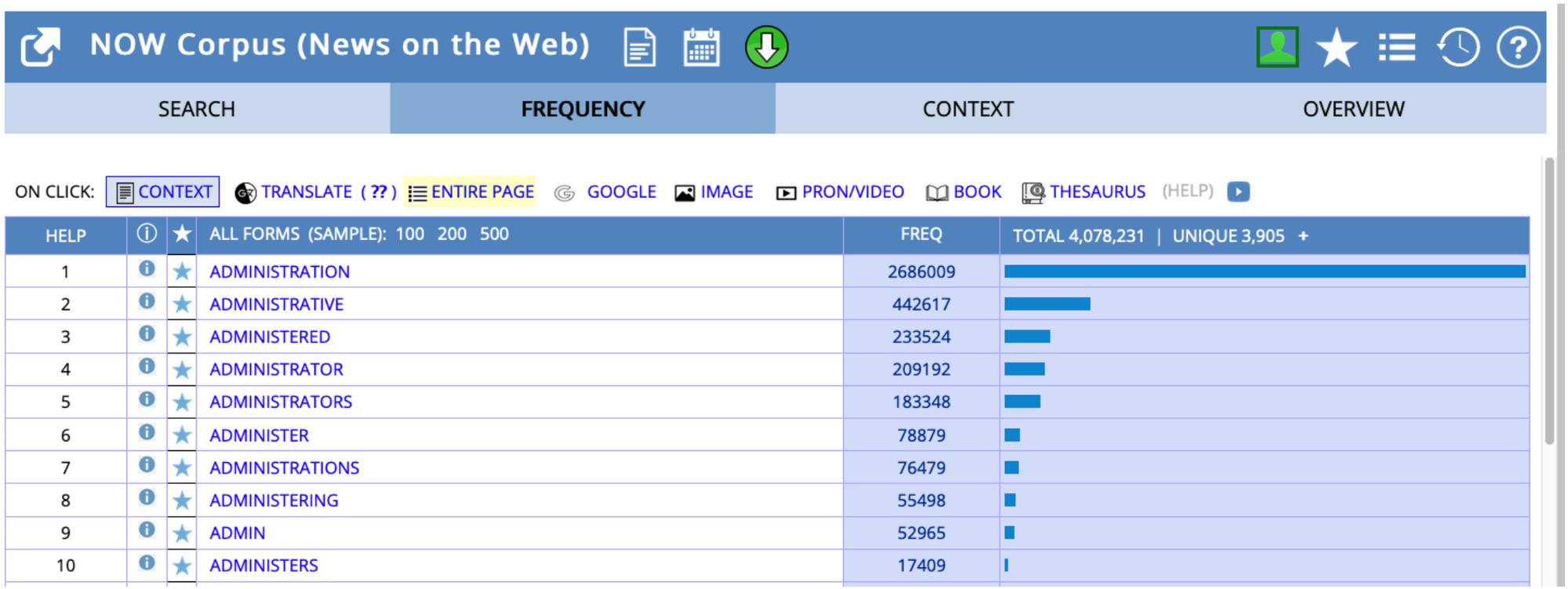


Chart View

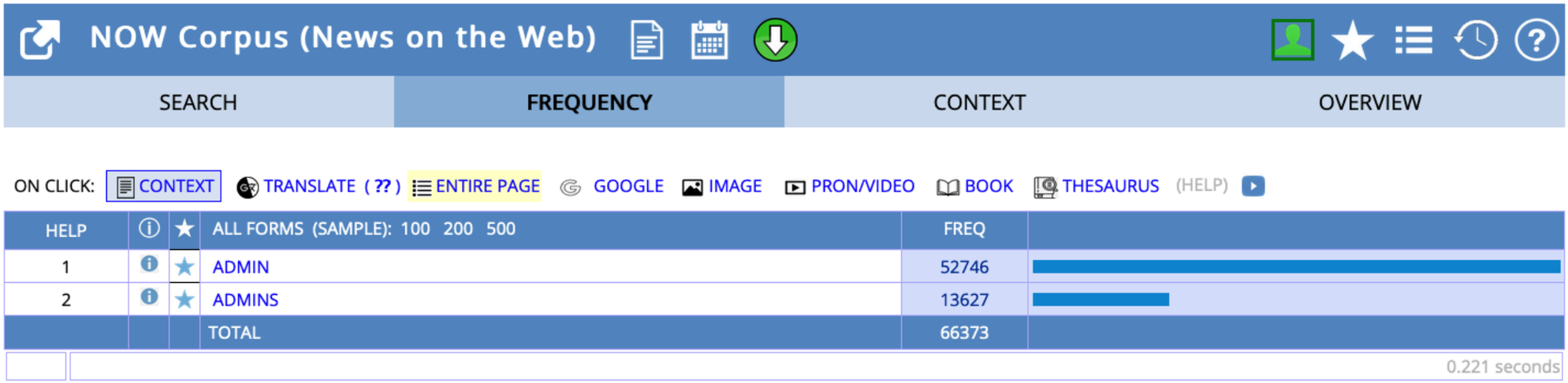
CLICK TO SEE CONTEXT See frequency by country

SECTION	ALL	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
FREQ	1889	47	56	73	73	67	64	104	143	117	187	259	274	304	121
WORDS (M)	17500	244.1	304.8	371.3	401.5	429.4	512.5	1,531.3	1,746.5	1,569.1	1,987.5	2,607.8	2,449.2	2,588.0	1,020.9
PER MIL	0.11	0.19	0.18	0.20	0.18	0.16	0.12	0.07	0.08	0.07	0.09	0.10	0.11	0.12	0.12
SEE ALL SUB-SECTIONS AT ONCE															

Query `decaf` in the NOW corpus

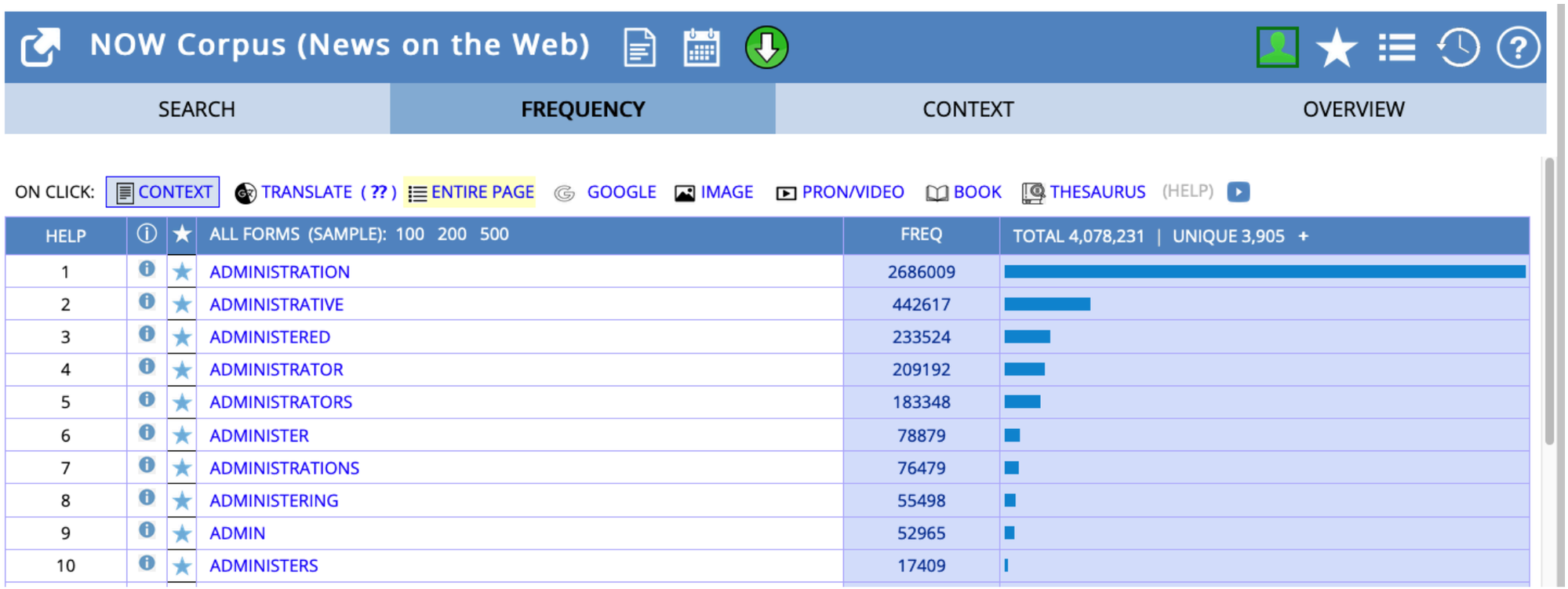
Query Syntax

Lexemes



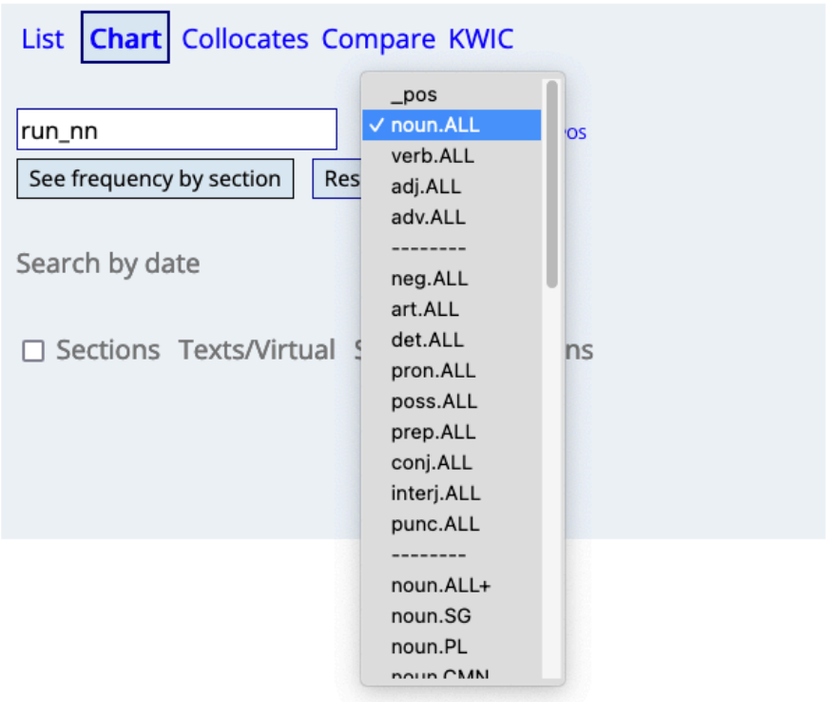
Query `ADMIN` in the NOW corpus

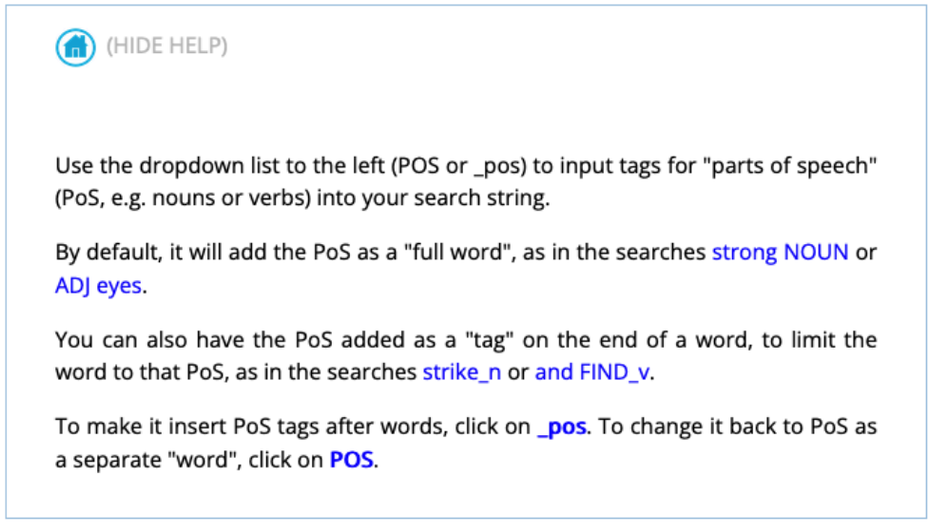
Wildcards



Query `admin*` in the NOW corpus (list view)

Word Classes





Query run_nn in the NOW corpus

NOW Corpus (News on the Web)

SEARCH | CHART | **CONTEXT** | OVERVIEW

SECTION: 2010 **(44,467)**
FIND SAMPLE: 100 200 500 1000
PAGE: << < 1 / 445 > >>

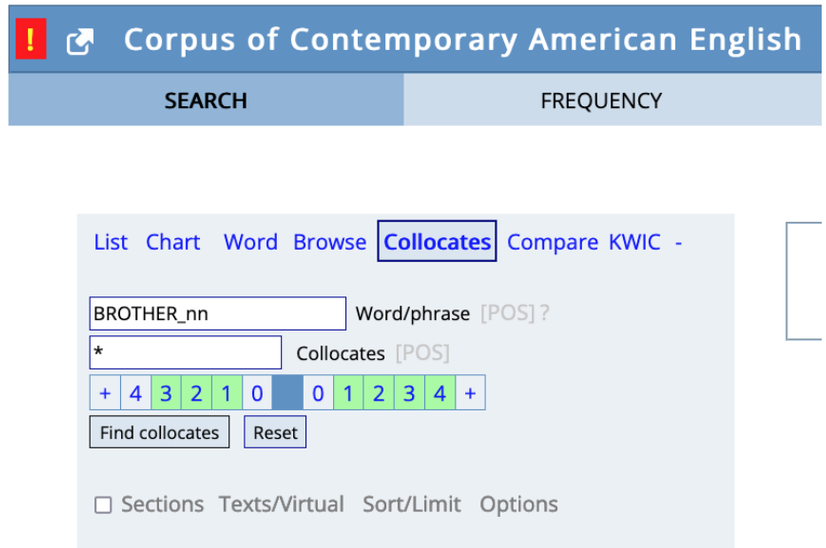
CLICK FOR MORE CONTEXT | SAVE | TRANSLATE | ANALYZE | HELP

1	10-12-31 JM	Jamaica Observer	but had resigned because he did not approve of the way the sport was being **run**. # He added that the new board had inherited a lot of debt
2	10-12-31 JM	Jamaica Observer	the bedmate? When an MP's bedmate buys land where the new road will **run**, pads to consultant in the firm about to get a state contract, ca
3	10-12-31 IN	Rediff	energy. # " And a dream of Ratan Tata is eventually cars will be **run** on water. My dream is also that, " he said, referring to
4	10-12-31 IN	Rediff	water-powered cars! # Tata Group chairman Ratan Tata's dream is to see cars **run** on water and he has invested $15 million in a start-up firm
5	10-12-31 IN	NDTV.com	empty and undamaged bottles of foreign-made liquor with the labels intact and used them to **run** the illegal racket. # The trio from Kerala u
6	10-12-31 IN	NDTV.com	" Both sides are very keen that the next session (of Parliament) should **run** in order... I am very, very optimistic on how the events will unfold
7	10-12-31 IN	Deccan Herald	temperature does not go up much, " Dr Charkoudian said, but if you **run** hard for an hour or so, you can have what seems like a fever
8	10-12-31 IN	Times of India	Jumbo Multi Axle' buses. These buses will have additional seating capacity and will **run** at a faster speed. The buses will be introduced on rou
9	10-12-31 IN	Times of India	of cost to the MSRTC for a period of one month. These will be **run** on the Swargate-Dadar route. Trips will leave daily at 7.30 am and 4 pm
10	10-12-31 IN	Moneylife Personal Finance Magazine	Katara couldn't get away because of alert citizen pressure. The drunken hit and **run** at Mumbai and Delhi are still hanging fire. Nothing seem

Query run_vv in the NOW corpus

Collocations

Collocates for BROTHER_nn



Collocates for BROTHER_nn

Corpus of Contemporary American English

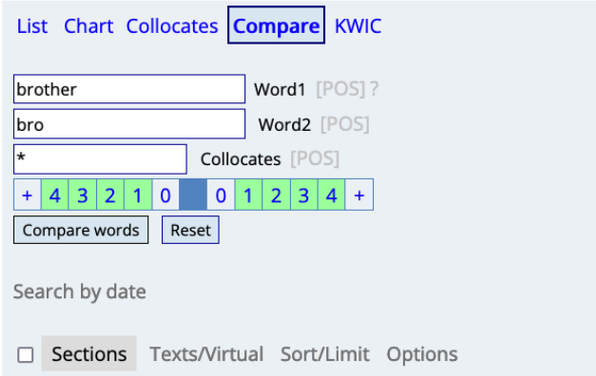
SEARCH | FREQUENCY | CONTEXT | OVERVIEW

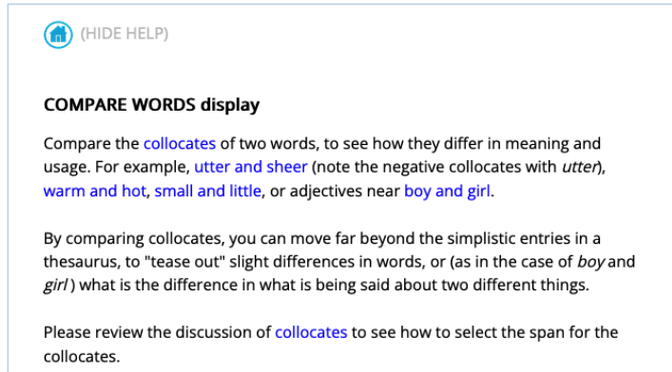
ON CLICK: CONTEXT TRANSLATE (??) ENTIRE PAGE GOOGLE IMAGE PRON/VIDEO BOOK THESAURUS (HELP)

HELP	RE-USE WORDS	FREQ	ALL	%	MI
1	SISTERS	6199	27129	22.85	7.60
2	OLDER	5388	94866	5.68	5.59
3	YOUNGER	4136	56091	7.37	5.97
4	SISTER	4042	91361	4.42	5.23
5	TWIN	1220	16630	7.34	5.96
6	LEHMAN	911	3654	24.93	7.72
7	KOCH	835	4685	17.82	7.24
8	WARNER	755	13555	5.57	5.56
9	ELDER	662	9975	6.64	5.81
10	MARX	336	5346	6.29	5.74

Collocates for BROTHER_nn

Comparing Collocates: *brother* vs *bro*





Comparing collocates: *brother* vs *bro*

Corpus Resources

COHA (Corpus of Historical American English)

- 400+ million words, 1810–2009
- decade-by-decade analysis possible
- fiction and non-fiction texts

COCA (Corpus of Contemporary American English)

- 1+ billion words, 1990–present
- text type categories: spoken, fiction, magazine, newspaper, academic
- enables register analysis

Step 1: Frequency Analysis in COHA

Target decades: `1850`, `1900`, `1950`, `2000`

CQL queries for modals:

Data Collection

- absolute and relative frequencies per decade
- model Excel sheet: https://1drv.ms/x/c/9a2ec97d593520f9/EezC1WmhjPNEiVR-eERIIU8BdRV5kbqEGw-17MMMJAr2gQ

Target Format

Lexeme	Type	Period	FreqAbs	FreqRel
would	core	1850	44567	2,695.15
would	core	1900	60305	2,743.97
would	core	1950	85122	2,969.93
would	core	2000	85403	2,452.57
may	core	1850	25536	1,544.27
may	core	1900	26706	1,215.17
may	core	1950	24891	868.45
may	core	2000	20047	575.7

Step 2: Text Type Analysis in COCA

CQL queries for modals:

- **core**:
  - → `can_v _v`
  - → `will_v _v`
  - → `may_v _v`
  - → `shall_v _v`
  - → `must_v _v`
- **peripheral**:
  - → `BE going to _v`

Text type categories:

- `BLOG`: blogs
- `WEB`: web pages
- `TV/M`: TV and movies
- `SPOK`: spoken
- `FIC`: fiction
- `MAG`: magazines
- `NEWS`: news
- `ACAD`: academic

Data Format

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Lexeme	Type	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	Average	StandDev	CoefVar
2	**would**	core	1,056.31	1,000.86	594.31	1,076.83	1,215.72	806.82	928.19	696.04	921.885	209.1633481	**22.6886594**
3	**may**	core	256.33	413.3	97.67	160.21	65.27	361.3	199.19	557.07	263.7925	168.9624943	**64.0512882**
4	**should**	core	1197.55	1188.85	1218.3	937.17	630.99	701.33	643.66	863.77	922.7025	253.377909	**27.4604121**
5	**must**	core	178.1	255.76	198.17	91.58	153.11	157.81	156.44	258.58	181.19375	55.83827513	**30.8168881**
6	**shall**	core	25.36	137.61	39.31	7.29	24.18	8.49	5.09	27.07	34.3	43.39451479	**126.51462**
7	**got to**	periphery	37.76	35.65	317.79	239.39	71.91	33.67	52.74	5.03	99.2425	114.2139167	**115.085691**
8	**going to**	periphery	265.68	213.6	441.95	1,169.22	269.52	142.37	219.08	29.94	343.92	353.3838462	**102.751758**
9	**have to**	periphery	539.38	701.22	1,097.83	973.66	674.51	424.58	458.04	142.59	626.47625	307.4013916	**49.0683233**
10	**need to**	periphery	482.84	417.52	542.92	300.76	228.03	274.19	225.53	271.48	342.90875	121.7302883	**35.499324**

Data format for text type analysis

Step 3: Statistical Analysis

Coefficient of Variation (CV)

Definition: Statistical measure describing relative variability of data

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100$$
$$= \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

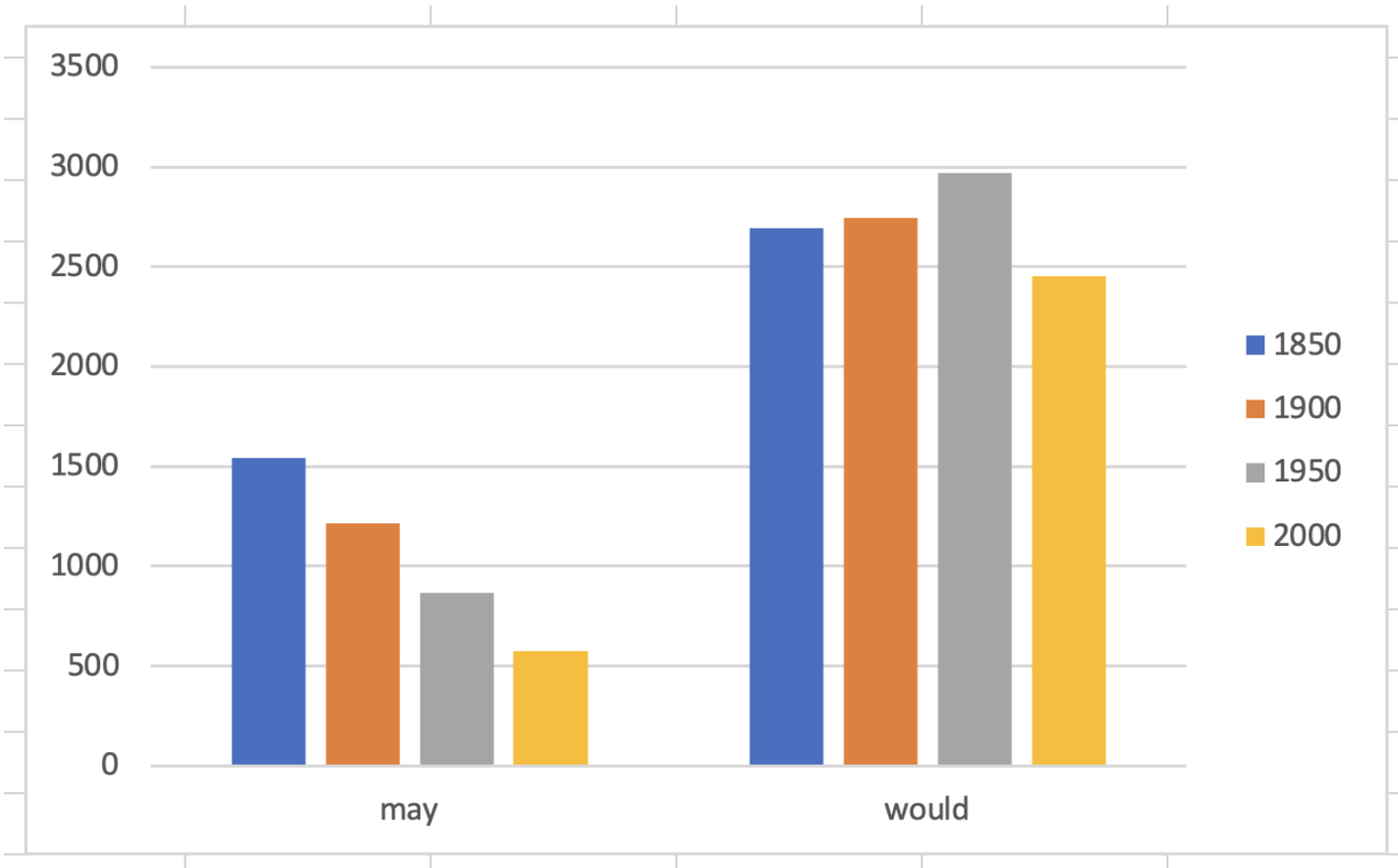
Excel Calculation

- Column 1: **Mean**: `=AVERAGE(range)`
- Column 2: **Standard Deviation**: `=STDEV.S(range)`
- Column 3: **CV**: `=(STDEV/MEAN)*100`

Expected Results

Frequency Changes Over Time

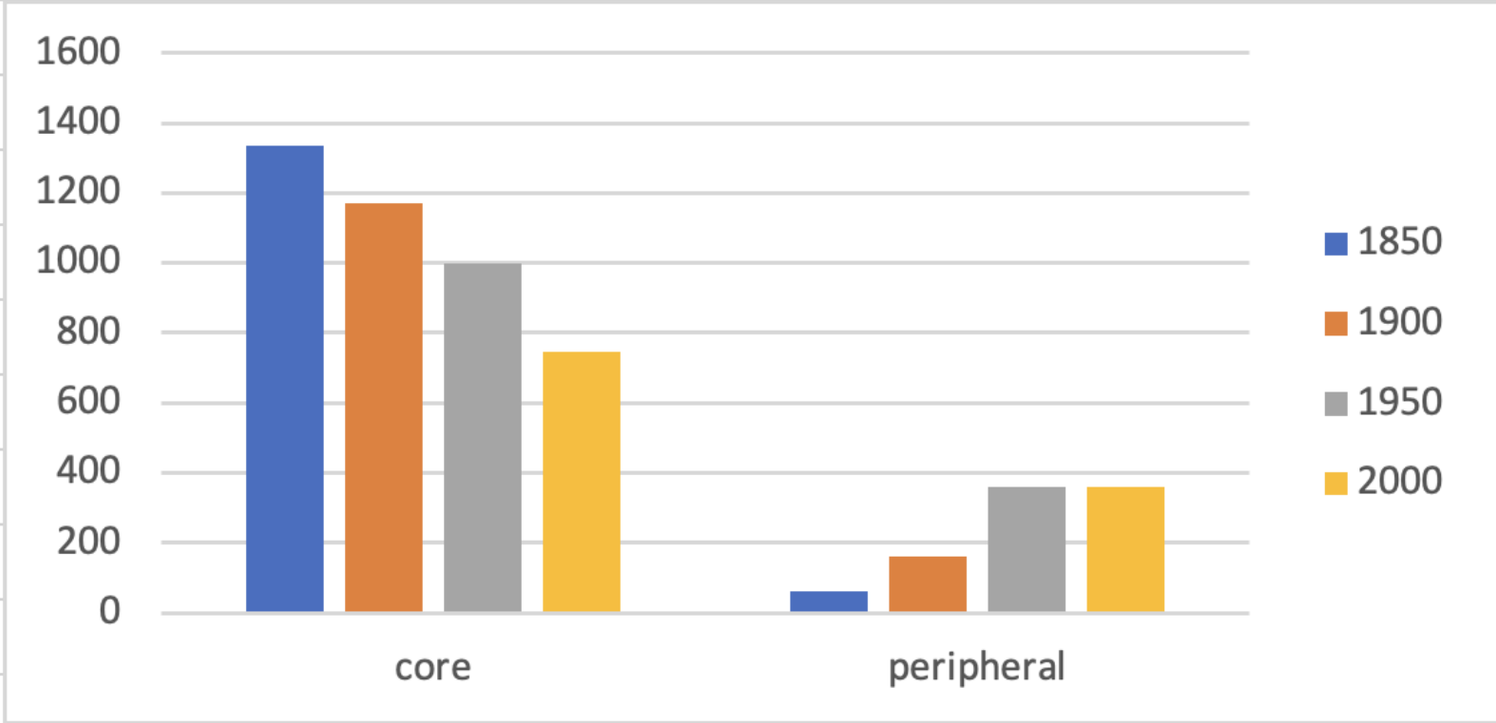
For individual modals



Frequency changes for *may* and *would*.

Across modal types

Average of FreqRel	Column Labels			
Row Labels	1850	1900	1950	2000
core	1334.765	1171.485	997.7383333	747.27
peripheral	61.98666667	161.44	358.7666667	361.04

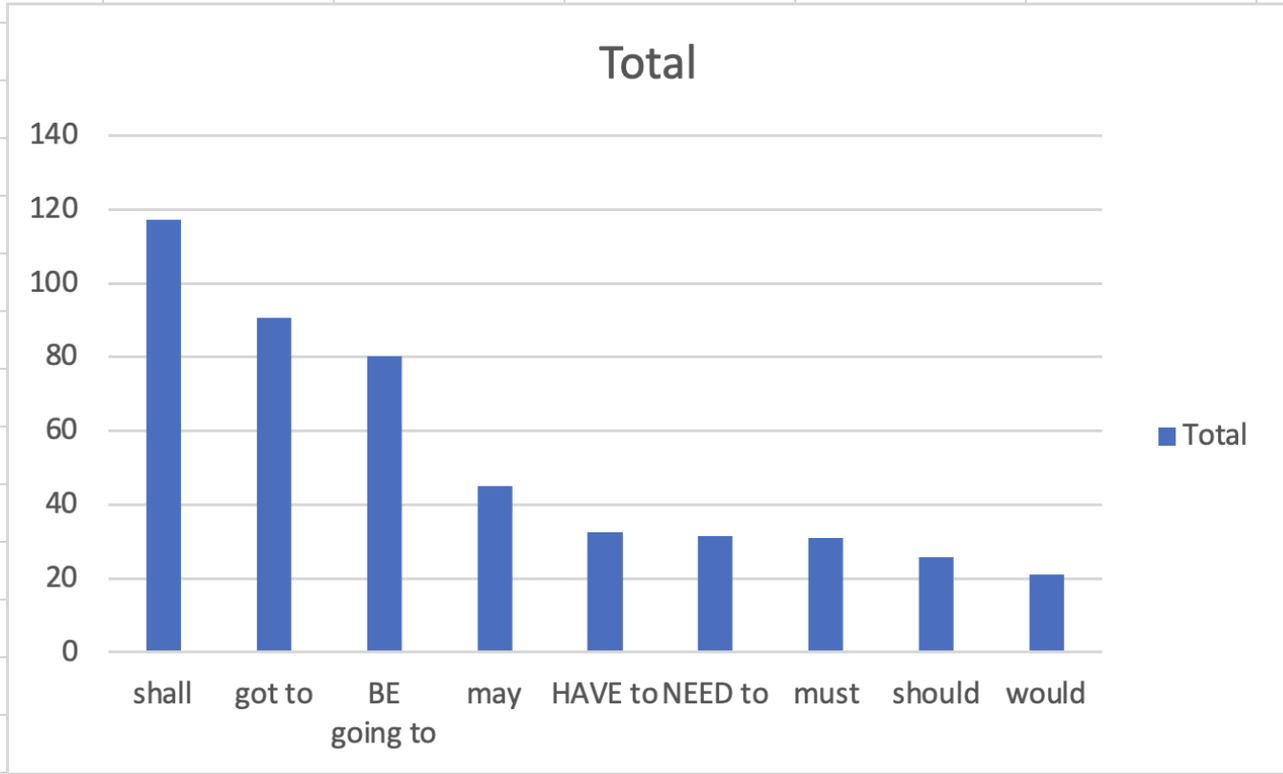


Frequency changes across modal types: core vs peripheral.

Text Type Variation

Text type variation for individual modals

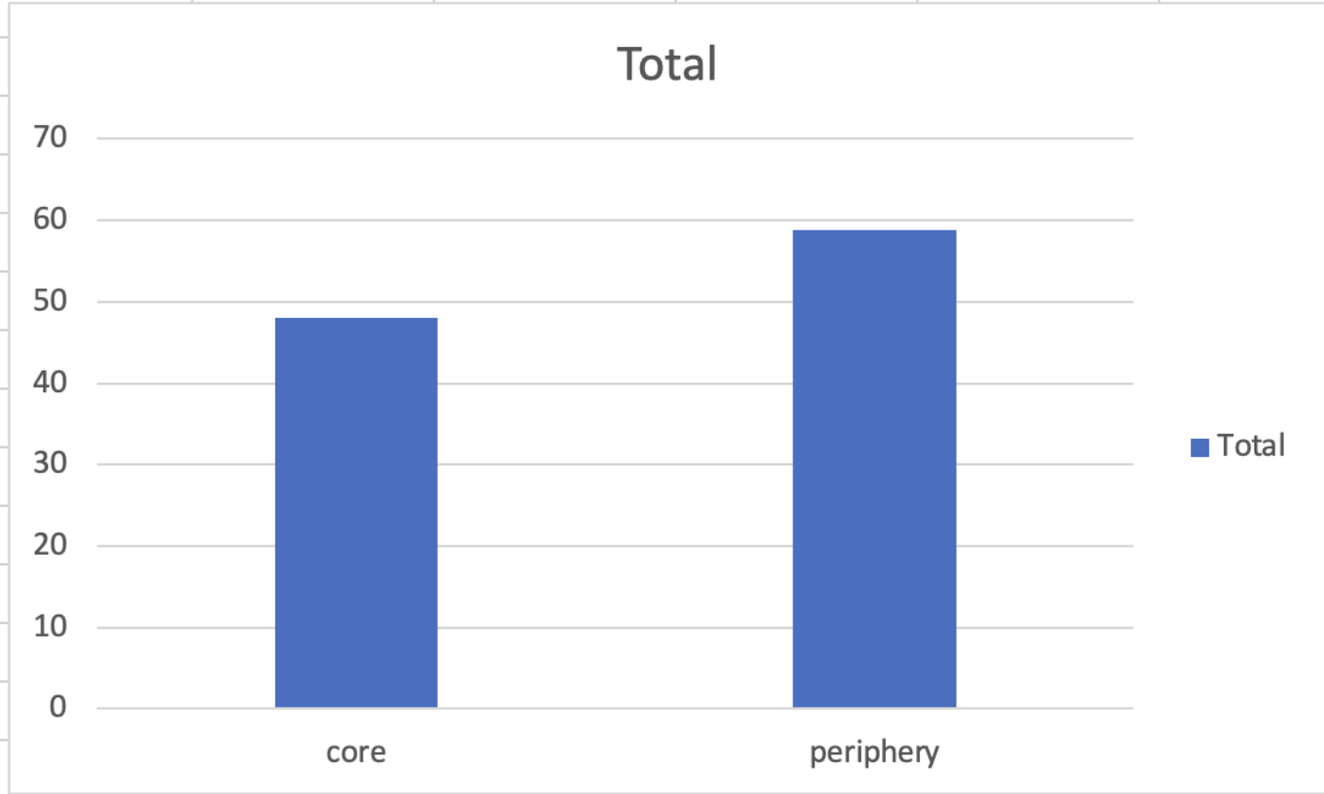
Row Labels	Sum of CV
shall	117.0671272
got to	90.65694482
BE going to	80.34030992
may	45.01880239
HAVE to	32.51406356
NEED to	31.42176765
must	30.92013087
should	25.68686345
would	21.09143731
Grand Total	**474.7174472**



Coefficient of variation for individual modals.

Text type variation for groups of modals

Row Labels	Average of CV
core	47.95687225
periphery	58.73327149



Coefficient of variation for groups of modals: core vs peripheral.

Study questions

1. Which modal verbs show the strongest frequency changes?
    - Are there differences between core and peripheral modals?
2. Which modals show the highest text type variation?
    - Are there differences between core and peripheral modals?
3. How do frequency changes relate to text type preferences?

Summary

- modal verbs show systematic frequency changes
  - → several core modals are declining
  - → peripheral modals are rising
- language change and text type variation seem to be related (in the case of modals)
- corpus methods provide empirical evidence

References

Hilpert, Martin. 2015. “Grammaticalization and the English System of Modal Verbs.” *Language Sciences* 47: 53–68.