Abstract

This paper uses a combination of natural language processing, a subfield of artificial intelligence/computer science, and traditional corpus linguistic techniques to explore how the language surrounding LGBTQ evolves through 20th and 21st century United States. I also provide historical and legal context to investigate if significant shifts in language surrounding LGBTQ people align with relevant historical events. The main analysis involves tracing the semantic shift of the word "gay" through the Corpus of Historical American English overtime. Using the methods k-nearest neighbors, collocation, stereotype quantification from word embeddings, and concordance, I found that the language surround the word "gay" has significant shifts in the years 1960 and 1990. These two years correspond with significant events in the gay rights movement and AIDS epidemic. Discussion is supplemented with analysis of other words relating to the LGBTQ community ("homosexual," "lesbian," and "queer") using similar methods.

1 Introduction

The LGBTQ+ community has faced discrimination in the United States since its founding. Following the American revolution, crimes such as "sodomy" were considered capital offenses. In the mid 1800s, 34 cities passed prohibitions against cross-dressing that allowed police to enforce gender norms and specifically target LGBTQ people. These laws remained in effect until as late as 1974. In 1952, the American Psychiatric Association's diagnostic manual lists homosexuality as a sociopathic personality disturbance that could be treated. In the same year, the Immigration Act barred "homosexuals and sex perverts" from enterring the United States. It wasn't until 2015 that the Supreme Court narrowly decided that the fundamental right to marry is guaranteed to same-sex couples.

Along with these legal discriminations, LGBTQ people have faced rampant homophobia in general society. When the AIDS epidemic began in 1981, it was commonly referred to as GRID (Gay-Related Autoimmune Disease) or as the "gay plague" as gay men were one of the primary groups afflicted. The name is an early indication of the perpetual associations between homosexuality and AIDS. For several years after the Center for Disease Control first discovered AIDS, the American government did little to address the epidemic since this disease mainly affected gay men, intravenous drug users, immigrants and racial minorities. At the same time, anti-gay reaction flourished across America with the election of Moral Majority ally Ronald Reagan. Thus, the growing medical crisis was met with public apathy and government inaction. It took until 4 years and over 12,000 American deaths before Reagan even publicly mentioned AIDS.

Homophobic sentiment continues to have lethal consequences in recent years. On June 12, 2016, a gunman attacked Orlando's Pulse dance club killing 49 people and wounding 50. Pulse dance club had established itself as one of central Florida's most vibrant centers for LGBTQ social life. This event marks the deadliest single incident targeting the LGBTQ community in U.S. history. In 2021, at least 57 transgender or gender non-conforming people were murdered. These crimes often go unreported, so there are likely more undiscovered victims.The majority of these people were Black and Latinx transgender women. Evidently, homophobia is ingrained in the history of the United States.

This paper uses a combination of natural language processing, a subfield of artificial intelligence/computer science, and traditional corpus linguistic techniques to explore how the languagae surrounding LGBTQ people change in the 20th and 21st century United States. I also provide historical and legal context to investigate if significant shifts in language surrounding LGBTQ people align with relevant historical events, such as the AIDS epidemic and the passing of various homophobic laws.

Natural language processing (NLP) relies on machine learning (ML) techniques. ML models can pick up on patterns in large data (e.g. a corpus) that are undetectable with human intuition. However, machine learning is a "blackbox," meaning researchers don't know *how* the model forms its conclusions. In other words, ML models find patterns in data without explaining what those patterns are. Thus, when using NLP for research, it's imperative to pair it with more traditional corpus linguistic research techniques to interpret the findings of ML models. Otherwise, analysis remains void of the nuance rooted in the actual text of the corpus.

This research involves analyzing the Corpus of Historical American English by decade for the years 1810-2009. Although all decades in this time frame are considered, relevant findings to the evolution of stereotypes against LGBTQ peoples lie mostly in the decades following 1960. I specifically focus on how semantic associations of the keywords "gay," "lesbian," "homosexual," "bisexual," "transgender," and "queer" change over time. Semantic association is determined using a myriad of quantitative and qualitative methods. Most notably, I use word-embeddings, which is a NLP technique that quantifies language, to statistically trace bias against this group over time. Finally, I contextualize the findings in history, demonstrating how the evolution of stereotypes align closely with relevant cultural shifts in the United States.

1.1 Overview of Previous NLP Research

Word embeddings are a machine learning framework that represents each word in a corpus as a vector. They capture meaningful semantic relationships between words through the distance between vectors since the vectors of similar words occur close to each other. For example, in the word embedding model Collober et. al. trained, the word "reddish" occured close to the word "pinkish" and the word "Jesus" occured close to the word "Christ" (1). Words that exist near each other are called "nearest neighbors," and indicate a close semantic relationship. Word embeddings also encode global relationships between words, allowing for the quantification of analogies. Mikolov et. al. demonstrate how the male/female relationship is automatically learned. In their paper, the vector representation of "King - Man + Woman" yields a vector very close to the vector representation of "Queen" (2). Using these two properties of word embeddings, Bolukbasi et. al. illustrate how they reflect cultural stereotypes, specifically regarding gender (3).

Hamilton et. al. develop a methodology for using word embeddings to quantify semantic change. They show that word embeddings are a promising tool for tracking the evolution of language across time (4). Some research already uses word embeddings to track historical semantic shifts in LGBT labeling words in the United States. Shi et. al. found that (5). However, this research only

analyzes the K-nearest neighbors (Section 2.2.1) of such words. This study expands on their methodology using more sophisticated statistical analysis. Specifically, Garg et. al. demonstrates that word embeddings can be used as a powerful tool to quantify historical trends and social change. They then developed metrics based on word embeddings to characterize how gender stereotypes and attitudes toward ethnic minorities in the United States evolved during the 20th and 21st centuries (6). I adapted their method to track how LGBTQ stereotypes evolved in the United States during the years 1810-2009.

1.2 Overview of Previous LGBTQ+ Stereotype Research

Current literature on homophobia and LGBTQ stereotypes tend to focus on modern discrimination. Raley et. al. (7) analyze prime-time network television during the fall of 2001 to identify the representations of gay male, lesbian, and bisexual characters. They found that gay men and lesbian representation serves to sometimes ridicule the community, but also to occasionally respect and regularize it. Tilcsik establishes that there is a significant discrimination against applicants who appeared to be gay in the job market, and that employers who lauded stereotypically male heterosexual traits are espeically likely to discriminate against openly gay men in the workplace (8). Klysing et. al. shows that sexual minorities are often stereotyped as their gender inversion. For example, homosexual women are seen as more similar to men in general than to women in general, whereas homosexual men are seen as more similar to women in general than to men in general (9).

There is some research on the history of homophobia. Tropiano also analyzes how gays and lesbians have been portrayed on entertainment television. He provides a historic outline of how television reflects major gay social issues writing “In the 1970s, the issue du jour was gay teachers. In the 1980s and 1990s, it was AIDS” (10). Wickberg also writes on how the evolution of homophobia directly embodies the United States’ cultural sensibility (11).

Other research establishes the connection between language and LGBTQ equality. Tavitas et. al. study the effects of gender pronoun use on mass judgments of gender equality and tolerance towards LGBT communities. They found that the use of gender neutral pronouns led to individuals expressing less bias in favor of traditional gender roles and categories and more favorable attitudes toward women and LGBT individuals (12).

2 Description of Data and Methods

2.1 Data

The data being analyzed consists of the Corpus of Historical American English (COHA). It is the largest structured corpus of historical English containing more than 475 million words from the 1820s-2010s. It includes more than 100,000 texts from fiction, popular magazines, newspapers, and non-fiction books, with the same genre balance decade by decade for 1810s-2009s. This research exclusively studies the lemmatized corpus, which reduces each word to its canonical form (e.g. walking

→ walk). Lemmatization reduces noise in statistical analysis since all derived forms of a word get treated as a single unit.

To track semantic and stereotype shifts over time, I partitioned COHA by decade. Each subsequent methodology is performed on every decade. Comparing results from different decades allows for diachronic analysis.

Some methodologies involve using word-embeddings extracted from COHA. Hamilton et. al. (4) released pre-trained word2vec historical word vectors for the lemmatized COHA partitioned by decade, which I downloaded for analysis instead of training new embeddings on the same dataset.

For every decade, Hamilton et. al. provides word embeddings for the top 50,000 words. Each word embedding is a 300-dimensional vector. However, words that occur infrequently do not carry enough semantic weight to build magnitude during the training process. In other words, some word vectors in the dataset are the 0 vector and thus become irrelevant for subsequent analysis. I cleaned the data of these useless vectors leaving the following number of word vectors for each decade:

1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
1321	4413	6908	7527	7449	7802	8079	8430	8510	9063	9161	10192	9892	9801	10043	9977	10065	10784	11521	12065

This reduction of the data demonstrated how many LGBTQ related words in various decades are too infrequent to garner statistical analysis from their word embeddings. For example, in the years 1810-1980, the word vector for “lesbian” is the mere 0 vector, and is thus unable to be analyzed in these decades. Further discussion on this continues in Section 3: Results and Analysis.

2.2 Methodologies

I use four distinct methodologies: (1) K-nearest neighbors, (2) collocation, (3) stereotype quantification, and (4) concordance. Each method is applied to a specific keyword (e.g. “lesbian”) for every decade 1810-2009 that has adequate data.

2.2.1 K-Nearest Neighbors Methodology

Word embeddings capture meaningful semantic relationships through their relative distances. In other words, if the distance between two word vectors is small, then they have high semantic association. The K-nearest neighbors algorithm finds the k closest word vectors (where k is an integer) to a given keyword vector. These neighboring words elucidate the associated meaning of the keyword.

I calculated and analyzed the 30-nearest neighbors of a given keyword for each decade and traced the shift of these nearest neighbors over time.

Although cosine similarity is often used as the distance metric in NLP research, Garg et. al. (6) opted for euclidean distance in their paper that analyzes the same word embeddings with a comparable method. Thus, I decided to use euclidean distance in my calculations of nearest neighbors.

2.2.2 Collocation Analysis

Collocation investigates which words co-occur with a given keyword. Using a window of ±4, the results rank the top collocates of the entire COHA before breaking down each collocate's frequency by decade.

Figure 1 - Example of collocation data

HELP	WORDS	ALL	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010	ALL	%
1	[LESBIAN]	297						1									1	1	11	78	95	110	1460	20.34
2	[GAY]	260	2	5	6	7	16	7	12	9	13	2	6	22	25	6	14	8	8	18	35	39	19396	1.34
3	[MARRIAGE]	223			1		1	1					1					1	1	12	93	112	42471	0.53
4	[BRIGHT]	184	5	8	10	14	13	23	12	14	12	4	12	12	15	12	8	6	3			1	50387	0.37
5	[FLOWER]	170	6	14	10	19	11	19	17	9	14	10	13	7	10	6	1	3		1			51060	0.33
6	[GRAVE]	134	6	15	14	11	14	9	13	13	18	9	5	4	1	1					1		36082	0.37
7	[GALLANT]	103	8	12	13	7	10	10	9	10	6	2	10	4	1	1							5876	1.75
8	[LAUGHTER]	90			5	5	6	6	8	3	6	15	9	11	4	2	3	2	3	1		1	18292	0.49
9	[BISEXUAL]	86																	10	10	17	49	311	27.65
10	[THRONG]	77	1	12	4	7	5	3	12	4	12	8	4	3	2								7353	1.05

This chart ranks the top 10 collocates of the word "gay" in COHA. Each cell reports the frequency of a collocate in a given decade.

2.2.3 Stereotype Quantification

Garg et. al. outlines in the paper "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes" a method based on word embeddings to characterize how gender and ethnic stereotypes in the United States evolved since 1910 (6). I have adapted this method to characterize the evolution of LGBTQ stereotypes in the United States since 1820.

Word embeddings can be used to measure the strength of association (i.e. embedding bias) between neutral words and a group. For example, let's overview the steps for quantifying stereotypes against the word "gay." The euclidean distance between the word vector for "gay" and the word vector for a neutral word, such as an adjective, measures the embedding bias between these two words. I compute this distance for a list of adjectives provided by Garg et. al. (6). For comparison, I also compute the embedding distance between words that represent the general population and the same list of neutral adjectives. For this metric, take the average vector of a group of words that represent the general population (e.g. "he," "she," "person," ...) and compute the euclidean distance between this vector and an adjective in the neutral word list. The metric for embedding bias to a specific neutral word is the distance for "gay" minus the distance for the general population. This metric is calculated for all words in the adjective list. A positive bias indicates the adjective has a stronger association with the general population, and a negative bias indicates the adjective has a stronger association with "gay." This method is performed on all keywords for all decades.

With these biases, I perform two forms of analysis. The first involves qualitatively analyzing the top 10 adjectives associated with the general population, which elucidates what the keyword is stereotypically *not* associated with, and the top 10 adjectives associated with the keyword, which elucidates what the keyword *is* stereotypically associated with. Comparing these adjectives for a specific keyword across decades demonstrates how stereotypes against the keyword evolved over time.

The second form of analysis involves running a Pearson Correlation test between two decades for a specific keyword. When calculating correlation, each data point represents an adjective in the list

with the x-coordinate being the bias for one decade and the y-coordinate being the bias for the other decade. The higher the correlation, the more similar the stereotypes against the keyword are for the two decades. This method is performed on all keywords across all combinations of decades.

2.2.4 Concordance

Concordance allows a researcher to search a corpus for a specific word or phrase and output all examples of that sequence in the context of the original text. I provide relevant concordance lines to supplement discussion.

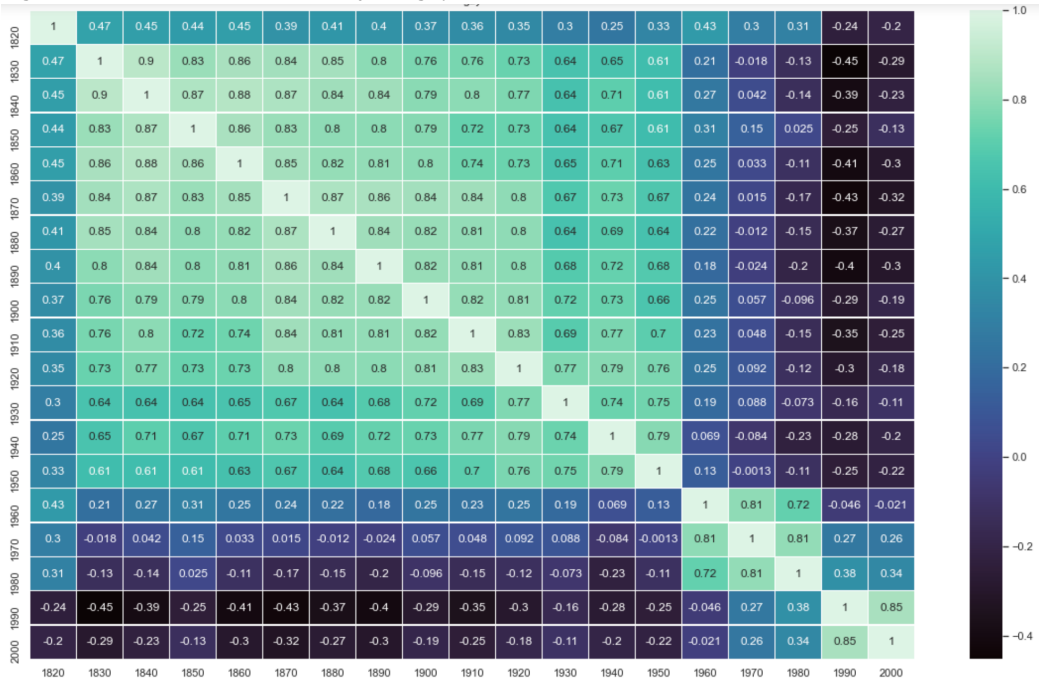
3 Results and Analysis

The bulk of this section involves a discussion on what the semantic shift of the word "gay" elucidates about stereotypes across 20th and 21st century United States. Additional discussions of the words "gay," "homosexual," "lesbian," and "queer" serve to provide a more nuanced analysis on attitudes towards the LGBTQ community as a whole.

3.1 Gay

I use the method described in Section 2.2.3 to trace how the description of the word "gay," through adjectives, has changed over time. Figure 2 shows the Pearson correlation in embedding bias scores for adjectives over time between COHA embeddings for each pair of decades. Note that there is not enough data to calculate the bias metric for the decade 1810 since the word "gay" occurred too infrequently in this partition of the corpus.

Figure 2: Pearson Correlations for "gay"



As expected, the highest correlation values are near the diagonals since attitudes are most similar to those from adjacent decades. More strikingly, the matrix exhibits three distinct blocks. There are distinct divides at 1960 and 1990, indicating that there was a significant shift in attitudes towards gay people at thse times.

3.1.1 Years 1820-1959

Prior to 1960, the word "gay" was used to mean carefree, cheerful, or bright. It was an adjective and did not refer to homosexual people.

Figure 3 displays the top collocates of the word "gay" with the frequency breakdown by decade.

Figure 3: Collocates of "Gay"

HELP	WORDS	ALL	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010
1	[LESBIAN]	297						1									1	1	11	78	95	110
2	[GAY]	260	2	5	6	7	16	7	12	9	13	2	6	22	25	6	14	8	8	18	35	39
3	[MARRIAGE]	223			1		1	1					1					1	1	12	93	112
4	[BRIGHT]	184	5	8	10	14	13	23	12	14	12	4	12	12	15	12	8	6	3			1
5	[FLOWER]	170	6	14	10	19	11	19	17	9	14	10	13	7	10	6	1	3		1		
6	[GRAVE]	134	6	15	14	11	14	9	13	13	18	9	5	4	1	1					1	
7	[GALLANT]	103	8	12	13	7	10	10	9	10	6	2	10	4	1	1						
8	[LAUGHTER]	90			5	5	6	6	8	3	6	15	9	11	4	2	3	2	3	1		1
9	[BISEXUAL]	86																	10	10	17	49
10	[THRONG]	77	1	12	4	7	5	3	12	4	12	8	4	3	2							
11	[BRILLIANT]	74	3	8	6	10	8	7	5	3	5	4	3	3	5	4						
12	[ACTIVIST]	67																15	14	18	12	8
13	[OPENLY]	64									1							1	1	6	24	31
14	[LIVELY]	64	3	9	7	4	2	8	5	2	4	2	3	4	1	3	4	1			1	1
15	[CHEERFUL]	64	2	6	5	6	5	5	6	2	5	6	6		4	2		2	2			
16	[ATTIRE]	55	2	4	13	7	3	9	2	5	4	2			1	1	1		1			
17	[CAVALIER]	53	1	8	12	5	12	7	1	4	1	1		1								
18	[MOOD]	53		2	1	2	4	5	4	3	2	2	5	3	7	5	3	3	1	1		
19	[LIGHT-HEARTED]	50	1	4	3	5	3	5	8	5	4	2	3	5	1			1				
20	[MERRY]	50	2	8	5	4	4	4	4	5	4			1	3	1	2		3			
21	[RIBBON]	48			3	7	3	3	4	5	4	3	2	2	3	2	2		1	1	2	1
22	[NINETIES]	45											1	4	10	9	4	8	5	4		
23	[PORTUGUESE]	44							44													
24	[HUMOR]	44		2	1	8	9	1	3	4	3	2	2	2	1	2		2		1		1
25	[CARELESS]	43	1	7	3	3	8	1	4	1	3	4	1	2	2		1				2	
26	[COSTUME]	42		3	6	4	2	4	4	2	1	4	1	1	5	2	2	1				
27	[TUNE]	42		1			3			4	3	3	6	5	5	5	2	4				1
28	[JOYOUS]	40		4	6	9	3	2	5	1	5	1	2				1	1				
29	[BLOSSOM]	40		4	9	4	4	2	3	3	1	3			4	2	1					
30	[FESTIVE]	38		2		3	6	6	2	5	1	5	2	1	3			1	1			

The most frequent collocations for the decades 1820-1950 include "bright," "flower," "grave," "gallant," and "laughter." These words align with the given definition of "gay" during this time period. The concordances from this time period also support this definition. For example:

> *1822* : "... the festive board, the free passage of the rosy hued goblet, and the **gay** and cheering discourse of social happiness, wore on the swift winged hours..."
>
> *1881*: "...the streets brilliant in the sun of an autumn day, into the town, **gay**, bustling, crowded, pulsing with vigorous life. How blue the sky was..."
>
> *1959*: "...flexible, they can be used for any mood or expression. White can be **gay** or somber, black can be stimulating or dreary, so I don't include..."

Evidently, the word “gay” is not in reference to the LGBTQ community for this time period and thus cannot elucidate any information on stereotypes. Section 3.2 explains that the word “homosexual” is used to describe this community during this time frame.

3.1.2 Years 1960-1989

Referring to Figure 3, the word “gay” begins to have collocates referencing the LGBTQ community, such as “lesbian” and “bisexual.” Other words that have connection to this community include “activist”, “openly” (as in “openly gay”), and “marriage.”

The nearest neighbors of “gay,” shown in Figure 4 include words that describe qualities and occupation.

Figure 4 - 30-nearest neighbors of “gay”

1960	1970	1980	1990	2000
gay	gay	gay	gay	gay
cheerful	jacques	burt	lesbian	lesbian
charming	elaine	promoter	activist	abortion
sentimental	charming	choreographer	abortion	activist
strangely	steven	seaman	minority	marriage
dazzle	cute	overweight	violence	woman
necklace	est	dis	jewish	minority
embroider	mystic	gentile	woman	conservative
mocking	wasp	aurora	feminist	openly
lively	militant	emerson	homosexual	civil
blonde	activist	werner	advocate	advocate
soprano	refinement	professionally	journalist	married
ornament	gifted	quartet	aids	homosexual
costume	whore	schoolteacher	racial	politician
flute	val	aristocrat	community	discrimination
husky	idealistic	enthusiast	sex	homosexuality
homely	barnum	entertainer	suicide	group
shiny	buffet	quip	hispanics	jewish
pleasing	gaily	disciplined	man	sex
soothing	lazy	sensational	discrimination	outspoken
sparkle	blush	lucas	black	religious
blouse	schoolteacher	visionary	ethnic	ethnic
dancing	peacock	il	citizen	atheist
khaki	ignorant	schroeder	ban	liberal
radiant	macdonald	filipino	christian	oppose
flannel	mint	festive	youth	feminist
gaunt	talented	constable	civil	religion
earring	melancholy	shaker	educate	muslim
plum	scissors	midwestern	commit	unmarried
frail	seethe	farce	hispanic	christian

The nearest neighbors surrounding “gay” likely include semantic associations from when “gay” was used as an adjective, such as “cheerful” and “charming.” Other words such as “activist,” “militant,” and “farce” likely serve to describe gay people. Further, the influx of occupation words, such as “schoolteacher,” “choreographer,” and “seaman,” are occupations that gay people are stereotyped to have. However, since machine learning is a black box, the algorithm does not explain the semantic reasoning for choosing these top neighbors. Thus, it’s better to analyze concordances for a fuller picture of the use of “gay” during this time.

Concordances show that gay people are being characterized as a community for the first time in mainstream United States. Still, throughout these decades, the word gay is still used as an adjective.

1965: " The child was going home, as she had come away from it, **gay**, irresponsible, and merry"

1979: " in her face whether she is sick or well, he feels her spirit as **gay** and playful as a girl's."

1980: "From far away came the sound of laughter and of a **gay**, bright, tinkling music."

The above are examples of the adjective use of "gay."

However, upon examining the concordances of "gay," the use of "gay" for describing gay people become more pervasive while the use of "gay" as an adjective becomes significantly less common.

1963: "It seems the only way I can do it is to mix myself up with blackmail and backstairs politics, and a bunch of **gay** boys and Lesbians."

1976: "Against bank rules. - I don't care. At least he's not **gay**. Plus, I love the way he talks. How about that suit?"

1988: "Jame is not really **gay**, you know, its just something he picked up in jail."

These above are examples of the "gay" refferring to gay people. Note that many of these concordances convey "gay" with negative connotations by associating them with criminal activity or defending that someone is not gay as if it were an insult.

The top adjectives in Figure 5 associated with gay elucidate what the Pearson's model picked up on in the semantic shift.

Figure 5 - Top 10 adjectives associated with "gay"

1960	1970	1980
noisy	idealistic	outgoing
sentimental	sophisticated	unreliable
polished	fashionable	irresponsible
unstable	noisy	intuitive
coarse	unstable	vulnerable
commonplace	informal	sage
cheerful	enthusiastic	cooperative
witty	humorous	humorous
homely	realistic	dominant
solemn	progressive	progressive

Throughout this time period, the adjectives become less obviously positive with words such as "cheerful" and "witty" making way for words such as "dominant" and "irresponsible." Further, as "gay" strays from its use as an adjective, gay-people become characterized as "noisy," "outgoing," and "fashionable," which elucidate some stereotypes about the community at this time.

Although the gay liberation movement began to gain traction throughout the 1950s in response to homophobic legislation and practices permeating the United States. The Lavendar Scare begain in 1950 based on the unfounded fear that gay men and lesbians "posed a threat to national

security because they were vulnerable to blackmail and were considered to have weak moral characters." Between 5,000 and 10,000 people lost their jobs as a result of the Lavender Scare, and the policy served to stigmitize the LGBTQ community on a national level. In 1958, the landmark case One, Inc. v. Olesen, the United States Supreme Court rules in favor of the First Amendment rights of the lesbian, gay, bisexual and transgender (LGBT) magazine "One: The Homosexual Magazine." It marks the first time the United States Supreme Court rules in favor of LGBTQ people.

The gay rights movement continued to flourish in the 1960s. In 1969, the Stonewall Uprising laid the foundations for the first pride parade in 1970. During Stonewall, LGBTQ people protested police who often harrassed a popular gay bar. This event is credited with reigniting the United States' modern LGBTQ rights movement.

The influx of using "gay" to refer to homosexual people starting in 1960 likely follows from this new wave of attention and activism. The changes in language surrounding "gay" in this corpus follow a decade after gay people began to garner a political audience at the national level. In other words, the shifting of language in response to historical events is not immediately solidified.

3.1.3 Years 1990-2009

The third block from Figure 6 represents a shift to politicization of LGBTQ people.

The adjectives most associated with the word "gay" for this time period reflect a community in protest with the government.

Figure 6 - Top 10 adjectives associated with "gay"

1990	2000
conservative	conservative
initiative	religious
radical	radical
religious	initiative
moderate	independent
hostile	moderate
traditional	aggressive
conventional	traditional
cooperative	outspoken
aggressive	dominant

The adjectives "radical," "initiative," "hostile," "aggressive," "dominant," and "outspoken" reflect a community that is seen as a threat to peaceful complacency.

Examining the nearest neighbors from Figure #, 1990 and 2000 see a greater number of words that adress political issues such as "activist," "abortion," "minority," "feminist," "racial," "discrimination," and "civil," "conservative," "liberal," and "politician." The concordances reflect this notion with political language surrounding the word.

> *1994:* "People were afraid of censure by **gay** activists, feminists, environmentalists -- now they are not because Rush takes them on"

2006: “while the " out " atheists tend largely to be older, male and white, their political views are not monolithic, said Brown and the other atheists gathered at Tommy's Joynt. They cover the range of opinions on everything from **gay** marriage to abortion to taxes to the war in Iraq” Gay people became a common topic in political conversation during these decades.

The AIDS epidemic began in 1981. The disease often infected gay men, and thus it was commonly referred to as Gay-Related Autoimmune Disease (GRID) which directly led to extreme stigmitazation of the disease. Many people believed only certain groups of people (i.e. gay men and drug users) could get the disease which led to irrationally negative attitudes towards people with AIDS. They essentially blamed the person with the disease for contracting it, which hindered the development of prevention and treatment strategies. Thus, people with AIDS not only struggled to find medical care and treatments, but also endured discrimination and social stigma.

In response to the epidemic and government inaction, the LGBTQ community began a new wave of activism forming organizations and protests. In 1982, Gay Men’s Health Crisis, the oldest AIDS service organization, formed in New York City. In 1987, San Francisco held the first and largest AIDS Walk. In the same year, activists founded the AIDS Coalition To Unleash Power, or ACT UP, in New York City. They organized many large scale protests including a take over of Wall Street to protest the profiteering of pharmaceutical companies in spite of the gross neglect to produce proper treatment for the disease. In 1988, they seized the FDA and in 1989 they took over St Patrick’s Cathedral. All of these protests forced national attention. Activism continued through the 1990s.

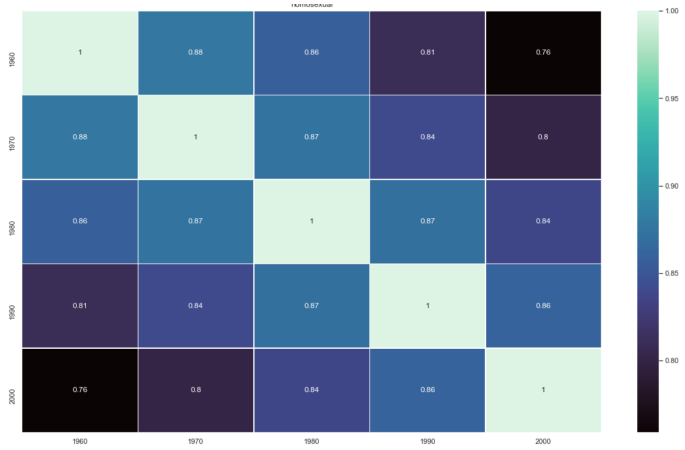
As the government finally responded to the epidemic in the late 90s, new political issues emboldened new protests. In 1996, the Defense of Marriage Act was signed into law defining marriage by the federal government as between a man and woman. Thus, states could use federal law to deny marriage equality. 2008 saw greater protests for marriage equality across the United States. In 2015, the Supreme Court made the landmark 5-4 decision in Obergefell v. Hodges to grant same-sex couples ithe right to full, equal recognition under the law.

Evidently, the LGBTQ community directly engages and communicates with the government in new ways starting in the 1980s. The shift to greater politicization of the LGBTQ community in 1990 follows from this wave of AIDS activism . Much like the shift in 1960, this change in language follows a decade after the major cultural shift that forced it. It takes time for societal language to respond to historical events that force a change in attitude.

3.2 Homosexual

The Pearson’s correlation test shown in Figure 7 reflects a gradual, uniform change in attitudes surrounding the word “homosexual” for the years 1960-2009. There is not adequate data to extract the word vector for prior decades.

Figure 7 - Pearson’s correlation for “homosexual”



Examining the collocates of "homosexual" in Figure 8 support the notion that 1960 marks a significant increase in the use of this word with top collocates including "bisexual" and "gay."

Figure 8 - Collocates of "homosexual"

HELP	WORDS	ALL	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
1	[ACT]	35														1	12	2	6	3	11
2	[HETEROSEXUAL]	30													2	1	3	3	8	4	5
3	[ACTIVITY]	30													2	2	3	7	6	6	3
4	[BEHAVIOR]	28													5	1	1	4		2	6
5	[EXPERIENCE]	24													3	3	1	3	11	2	1
6	[MALE]	22													2		2	6	3	3	3
7	[RELATIONSHIP]	18													1	1		2	4	3	3
8	[CONTACT]	16													1	1	8	1	5		
9	[SEX]	16														1		3	3	3	4
10	[BISEXUAL]	15																	6	6	1
11	[COUPLE]	15															2		5	3	3
12	[TENDENCY]	14										1		2	1	1	1	3	2	1	2
13	[HOMOSEXUAL]	13													1		1	6	5		
14	[DRUG]	13															1	1	9	1	
15	[ENGAGE]	13															3	4	2	1	1
16	[MARRIAGE]	13																2	4	1	5
17	[TEACHER]	13																9	2	1	1
18	[PROSTITUTE]	12													1	1	1	2	3	2	2
19	[FANTASY]	11														1	1	1	8		
20	[CONDUCT]	11																4		3	
21	[RAPE]	10															1	2	4	1	1
22	[GAY]	10																1	1	4	4
23	[ORIENTATION]	9															6			1	2
24	[COMMIT]	8												1			4		2		
25	[LATENT]	7													3		1	1	2		
26	[ABORTION]	7																2	3		1
27	[EXCLUSIVELY]	7														2		2		2	1
28	[OPENLY]	7																1	2		3
29	[SEXUAL]	7															2	1	2	1	
30	[ENCOUNTER]	7																1	3	1	1

However, the use of "homosexual" tends to focus on acts or experiences rather than a person.

1964: "He admitted committing at least four **homosexual** acts, some of them for pay"

1980: " I want to know about your so-called **homosexual** experiences. This is a very serious matter."

2011: "He still holds to what he terms " the very clear Biblical teaching " that **homosexual** behavior is not in God's design for sexuality and is sinful."

Thus, the *experiences* of LGBTQ people become more prominent at the same time the LGBTQ community begins to garner more attention in the public eye.

The influx of this word in 1960 likely follows from the same reasons dicussed in Section 3.1.2 since the second and third blocks of Figure 2 aligns exactly with the years the word "homosexual" began to solidify its presence in the corpus.

3.3 Lesbian

Since the only decades where "lesbian" occurs frequently enough to extract word embeddings are 1990 and 2000, a Pearson's correlation test is unhelpful in this analysis. Nonetheless, an examination of collocates illuminate an influx of the use "lesbian" in reguards to LGBTQ people at the same the word "gay" began garnering a more politicized connotation.

Examining the collocates of "lesbian" in Figure 9 show strong relations to the LGBTQ community starting in 1990 with "gay," "bisexual," and "activist" being top collocates for the decades 1990 and 2000.

Figure 9 - Collocates of "lesbian"

HELP	WORDS	ALL	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
1	[GAY]	260						1										1	8	70	82
2	[BISEXUAL]	62															1			13	14
3	[COMMUNITY]	32																1		17	6
4	[COUPLE]	28																1	1	5	6
5	[SEX]	18																2	1	3	8
6	[SCENE]	17											3					2		1	7
7	[SAPPHO]	15		1	1	1	11								1						
8	[LESBIAN]	14					1											1		4	
9	[IDENTIFY]	11																	1	3	1
10	[ACTIVIST]	10																		4	2
11	[LOVER]	10																		2	4
12	[RELATIONSHIP]	10														1		2	1	3	1
13	[WEDDING]	9																		8	1
14	[ORGANIZATION]	9																		1	2
15	[SEAGULL]	8																		8	
16	[MOM]	8																		1	4
17	[FILM]	8																1		1	4
18	[ARTIST]	8																	1	7	
19	[FEMINIST]	7																	1	3	1
20	[ARCHIVE]	7																			

When appearing next to "gay" and "bisexual," it's often in reference to the larger LGBTQ community as a whole.

1993: "A chapter of the national Gay, **Lesbian**, Bisexual Veterans of America is forming in Utah to assist members of the military, whether active or retired, who are homosexual or bisexual."

The influx of this word in relation to the LGBTQ community in 1990 likely follows from the same reasons discussed in Section 3.1.3. The third block of Figure 2 aligns exactly with the years the word “lesbian” began to solidify its presence in the corpus.

3.4 Queer

The word “queer” was originally meant as an insult when referring to the LGBTQ community. It was, and still is, used as an adjective outside of this context. As evident from Figure 10, the top collocates have no obvious connection to the former use of the word.

Figure 10 - Collocates of “queer”

HELP	WORDS	ALL	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
1	[SORT]	197	2	2	5	7	5	14	15	14	17	29	31	27	17	6	3			2	
2	[QUEER]	168		2	4	22		3	7	4	12	20	29	19	12	7	5	3		1	4
3	[FELLOW]	103		6	6	6	4	6	13	11	8	14	12	5	4	2	4		1		
4	[MIGHTY]	73			1		4	5	8	7	9	16	9	6	4	2			1	1	
5	[NOTION]	71	1	2	4	3	3	8	5	5	3	10	10	7	4	2	3				
6	[NOISE]	58			3	2	1	1	2	2	5	14	11	5	5	1	3	1	1	1	
7	[FOLK]	54	1	1	1		3	2	5	2	12	5	5	6	4	2					4
8	[SENSATION]	50		3	1				3	1	3	11	6	8	1	6	4		3		
9	[CREATURE]	46				1	2	1	1	6	7	7	8	6	2	2	1		1		
10	[CHAP]	43	1	2	2	1	2	2	4	1	6	8	11	2	1						
11	[DUCK]	37						1		2	1	1	7	7	6	3	4		3		2
12	[TWIST]	33							1		5	6	12	2	4	3					
13	[MIXTURE]	30				1	1	5	4	3		1	3	6	4	1			1		
14	[DOING]	24		2			2	2	1	4	5		1	3	4						
15	[QUAINT]	21				2	2	3	3	2	2	3	2		2						
16	[OLD-FASHIONED]	20	1		1		1		1	2	1	6	3	2	1	1					
17	[AWFULLY]	19						1		3	1	2	4	3	3	1	1				
18	[SOMETHIN]	17					2	2			1	7	3	1				1			
19	[ANYHOW]	14		2				1	1	3	1	2	1	2			1				
20	[COSTUME]	14				1	2	1	2	1	2	1	2	2							
21	[DAMNED]	13									1	1	2	3	3		1	2			
22	[SPECIMEN]	12						1	2	1	1		2	2	2		1				
23	[FREAK]	11		1		1			4	1	2		2								
24	[LUMP]	11								1		4	4	2							
25	[MAMMA]	11				2		4	1		2	2									
26	[GAIT]	10			1						1		3	3		1	1				
27	[JUMBLE]	9						1	1		3	1	2	1							
28	[SMEAR]	9																		1	
29	[ACTIN]	8							1		4	1		1		1					
30	[JIMMIE]	8										5		3							

Word vectors cannot differentiate between these two uses of the word, thus making it an inadequate method of analysis.

A brief analysis using concordance and collocates does demonstrate the word “queer” to primarily refer to gay people starting in 1990.

Figure 11 - Collocates of “queer” 1990-2009

1990

1	[FUCKING]	4
2	[DADDY]	4
3	[BOOK]	4
4	[WORD]	4
5	[STYLE]	3
6	[ACT-UP]	2
7	[FAGGOT]	2
8	[SENSIBILITY]	2
9	[GAY]	2
10	[CONTENT]	2

2000

1	[EYE]	11
2	[STRAIGHT]	8
3	[QUEER]	4
4	[FOLK]	4
5	[VOICE]	4
6	[KIND]	4
7	[GAY]	3
8	[CAUSE]	3
9	[HOMOSEXUAL]	2
10	[DUCK]	2

Although there are examples of queer being used as a slur in the 1990s, the word loses this negative connotation by the 2000s. The following examples show "queer" as a slur in the 1990s:

> *1991:* " I don't remember. What do you think you're looking at? You fucking **queer**. Fucking scum. You think Nick gives a shitwhat people think? You knowwhat"
>
> *1995:* "You lying filthy fucking slut are gon na watch me help your little fairy boytoy shuffle off this mortal coil Kiss your puny, worthless cock goodbye Any last words, **queer**?"

In the same decade, there are also examples of it being used by gay people in reference to their community.

> *1996:* "I would even suggest, though I can't quickly say how he does it, that the way Henderson has written about terrain, landscape, and geography is bent, **queer**, gay original style. And once again, a **queer** author has taken a genre ... "
>
> *1997:* "My brother gave me the most extraordinary book. " **Queer** America - - from A to Zed. " Am I in it? It gives the names of all the gay men and lesbians in this country."

By 2009, the top collocates lack examples of queer being used as a slur.

> *2007:* "Yet Asim says individual groups must choose some level of consensus to allow others to use the words with impunity. " With **queer**, some homosexuals say it's OK and some say it's not. That's between homosexuals, " says Asim, a reporter for the Washington Post"

Thus, the corpus reflects the end of the transformation of this word from slur to reclaimed identifier.

The LGBTQ community began to reclaim the word "queer" in 1990 when an activist group dubbed "Queer Nation" was founded in response to the AIDS epidemic the same year. As described in Section 3.1.3, the 1980s-1990s saw an explosion of queer activist groups because of the AIDS epidemic. One of these groups is Act-Up, which is a top collocate of "queer" in the 1990s. Although "queer" has been used as a slur before this time, the language in the corpus only reflects its gay connotations after the community started reclaiming it. Thus, the corpus reflecting the use of queer to

describe LGBTQ people beginning in this decade is likely due to this reclamation of the word as well as the history dexribed in Section 3.1.3.

3.4 Bisexual and Transgender: A note on data

The infrequent occurances of the words “bisexual” and “transgender” make word vector analysis impossible. These two words are thus excluded from the study.

4 Summary and Conclusion

This analysis involved tracing the semantic shift of the word “gay” through COHA overtime. Using the methods k-nearest neighbors, collocation, stereotype quantification from word embeddings, and concordance, I found that the language surrounding the word “gay” has significant shifts in the years 1960 and 1990. These two years correspond with significant events in the gay liberation movement and AIDS epidemic.

Discussion is supplemented with analysis of other words relating to the LGBTQ community (“homosexual,” “lesbian,” and “queer”) using similar methods. An influx in the use of these words align temporally with a shift found in the word “gay.” Further, the reclamation of the word “queer” likely contributes to a greater representation in modern decades.

This paper establishes a connection between history and semantic shifts in language reguarding the LGBTQ community. **Changes in language often lag behind the events that trigger a cultural shift in attitudes, indicating that a societal change in language follows from political action. Political activism drives a change in majority attitude, rather than political change following from a change in societal attitude.**

Future work could involve analyzing the words “bisexual” and “transgender” using accessible methods along with slurs used against sexual minorities. For example, the word “queer” was originally used as an insult to LGBTQ people, but was later reclaimed by the community. A more indepth analysis of this transformation along with the slurs “dyke” and “fag” could elucidate insite into homophobia from different eras since these words carry negative connotations in reguards to LGBTQ people.

[Note: I changed the focus of this paper to focus on the shifts in language overtime rather than specific stereotypes found against LGBTQ people. I still have the data that could discuss what these stereotypes are, but I left it out of discussion since I found it muddled the main argument. I can include it if you think it would supplement analysis nicely, though.
Some of these observations include:

- A support for lesbians being masculanized and gay men being femminized in stereotypes
- lesbians being fetichized/sexualized
- homosexuals/gay men being sexualized]

References

1. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011, March). Natural language processing (almost) from scratch. https://doi.org/10.48550/arXiv.1103.0398
2. Mikolov, T., Yih, W., & Zweig, G. (2016, February). Linguistic regularities in continuous space word representations. Microsoft Research. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/rvecs.pdf
3. Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., & Kalai, A. (2016, July). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. https://doi.org/10.48550/arXiv.1607.06520
4. Hamilton, W., Leskovec, J., & Jurafsky, D. (2016, May). Diachronic word embeddings reveal statistical laws of semantic change. https://doi.org/10.48550/arXiv.1605.09096
5. Shi, Y., & Lei, L. (2020). The evolution of LGBT labeling words: Tracking 150 years of the interaction of semantics with social and cultural changes. English Today, 36(4), 33-39. doi:10.1017/S0266078419000270
6. Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, 115(16). https://doi.org/10.1073/pnas.1720347115
7. Raley, A., Lucas, J. (2006). Stereotype or Success?, Journal of Homosexuality, 51(2), 19-38. doi:10.1300/J082v51n02_02
8. Tilcsik, A. (2011). Pride and Prejudice: Employment Discrimination against Openly Gay Men in the United States. American Journal of Sociology, 117(2), 586–626. https://doi.org/10.1086/661653
9. Klysing, A., Lindqvist, A., & Björklund, F. (2021). Stereotype content at the intersection of gender and sexual orientation. Frontiers in Psychology, 12. https://doi.org/10.3389/fpsyg.2021.713839
10. Cleary, J. (2003). The prime time closet: A history of gays and lesbians on TV. Journalism and Mass Communication Quarterly, 80(4), 993-994. https://www.proquest.com/scholarly-journals/prime-time-closet-history-gays-lesbians-on-tv/docview/216931990/se-2?accountid=14953
11. Wickberg, D. (2000). Homophobia: On the Cultural History of an Idea. Critical Inquiry, 27(1), 42–57. http://www.jstor.org/stable/1344226
12. Tavits, M., & Pérez, E. O. (2019). Language influences mass opinion toward gender and LGBT equality. Proceedings of the National Academy of Sciences, 116(34), 16781–16786. https://doi.org/10.1073/pnas.1908156116