A Corpus-based Approach to Reclaiming *"Queer"*

Blanca Alonso Gonçalves
University of Groningen
b.alonso.goncalves@student.rug.nl

April 2025

1 Introduction

The Cambridge Dictionary of English (Cambridge University Press, n.d.) defines the adjective "queer" as *"having or relating to a gender identity or a sexuality that does not fit society's traditional ideas about gender or sexuality"*, and under this definition it includes a note that reads *"Queer can be offensive to some people. Only use this word if a person describes themselves in this way."* In fact, the use of "queer" as a neutral descriptor for LGBT people is relatively recent, as originally it was used only as an insult. The Merriam-Webster dictionary shows a similar note on the usage of "queer", included below:

The adjective queer is now most frequently applied with its meanings relating to sexual orientation and/or gender identity (...). When these meanings were developing in the early 20th century, they were strongly pejorative, echoing the negative connotations of the word's older meanings, which included "weird," "suspicious," and "unwell." But the adjective today is commonly used as a positive or neutral self-descriptor, and also has wide use as a neutral broad descriptor for a large and varied group of people. (...) *The term is also prominent as a neutral term in academic contexts that deal with gender and sexuality. Current neutral and positive uses notwithstanding, the word's long history of pejorative use continued into the current century, and some people still find the word offensive in any context.*

The question of whether the term "queer" is appropriate to refer to LGBT people has been a topic of debate since the 1980's (Gamson, 1995; Jacobs, 1998). Some people reclaimed the word that was used against them and tried to turn it into a neutral or positive term for the LGBT, while others refused to refer to themselves as such and found it insulting. Jacobs (1998) provides the following quote from the style guide *The "OUT!SPOKEN" Styleguide: A Guide for the Media on Lesbian, Bisexual, Gay and HIV/AIDS Issues (1994: 14)*, which mentioned the debate and advised against the use of the term:

Do not use. A term historically and currently used pejoratively towards lesbians, gays, and bisexuals. However, many people at whom the term was directed have reclaimed it and use it in a highly politicized sense. Among bisexuals, lesbians, and gays there are differing opinions as to whether this term should or should not be used. The term 'queer' may be used appropriately by the community and quoted as such, but because of the pejorative and threatening connotations of the word when used by heterosexuals it should not be used by those outside the community.

There have been studies regarding the use and perception of the word "queer", but most of them did not use computational or corpus-based approaches but instead relied on surveys or interviews. In a study by Zosky and Alberts (2016), college-aged attendees to the Midwest Bisexual Lesbian Gay Transgender Ally College Conference in the United States were interviewed on their opinion of the word "queer". This study found that 76.7% of the participants reacted positively and 5.6% reacted negatively to the use of "queer" as a term for self-identification, while 57.7% reacted positively and 16.3% reacted negatively to the use of "queer" to refer to others. They stated that these results are evidence of a positive shift in the perception of the word "queer", as older generations showed a much more negative reaction to the term, which is consistent with Jacobs (1998).

Zosky and Alberts (2016) also concluded that *"Perhaps the Queer Nation anthem cry "We're queer, we're here. Get used to it!" has been partially realized. The younger generation seems to have "gotten used" to the term queer."* However, this study covers only the perception of a specific sample of queer people and is not representative of the general sentiment associated to the word "queer" in other contexts. A computational and corpus-based approach to the question of a possible change over time of the sentiment attached to the word "queer" that uses a large collection

of linguistic data from different sources would give insights on how the general population uses the term, including what the intensity and polarity of the sentiment associated to "queer" are in different contexts and sources across many years.

Jacobs (1998) concluded that *"It remains to be seen, then, whether non pejorative 'queer' will achieve some measure of success as it moves from the originating speech community to the larger society."*, and this is exactly what this study aims to explore. Thus, our research question is whether the sentiment associated to the word "queer" by American English speakers has shifted positively between the years 1990 and 2019. Given the positive trend in the perception of the general population regarding queer rights, we expect this social change to be reflected in language use, resulting in changes in the sentiment of the contexts where the word "queer" is used, with positive sentiment increasing and negative sentiment decreasing over time.

As far as we know, the use of R tools for sentiment analysis of corpora has not been applied to study the use of the word "queer" yet, but similar methodologies have been used to explore the perception of other sociopolitical issues such as climate change (Mi & Zhan, 2023). We take inspiration from these studies as the sentiment analysis tools they used have proved to be successful.

2 Methods

All the data and code used in this study can be found in this Github repository.

2.1 Data

We use data from the Corpus of Contemporary American English (COCA), which consists of over 950 million words from 485,000 texts published between 1990 and 2019 and including fiction works, magazines, newspapers, academic publications, blogs, webpages, TV and film subtitles, and transcriptions of spoken speech (Davies, 2008-). We chose the COCA because it is a widely used source of data for corpus analysis that has proved to be useful in previous studies on the usage of "queer" and other related vocabulary (Motschenbacher, 2002; Hanneder, 2023), and because it encompasses thirty years of data, which allows for the study of the change in usage of a word over this time.

The full-text dataset is not available for download for free, and purchasing the corpus data was not a possibility for us. There are free downloadable samples on the COCA website, but these do not contain enough instances of the word "queer" to be useful for this project. The only option available for us then was to use the website interface of the corpus to search for the data necessary for our goal. The word "queer" was searched on the COCA using the list option of the website, which returned just over 5000 results. We were interested in the contexts in which the term "queer" appears, so we saved the concordances into five keywords-in-context (KWIC) lists. This was necessary because the website imposes a limit on the amount of concordance lines that can be expanded and explored at the same time. Then, each of the KWIC lists was copied into a CSV file using Excel, so they could be easily opened and processed in R. These CSV files together constituted the dataset used for our analysis. The five files were imported into R and combined into a single dataframe containing 5,006 observations across three variables: the year in which the text was published, the source of the text, and the text itself.

2.2 Data Exploration

Before going into the question of sentiment analysis, we explored the general usage of the word "queer" in the thirty-year period that the corpus comprises. The chart option of the COCA website gives us the bar chart in Figure 1, where we can already appreciate some trends. There is a clear increase in frequency across each five-year period, with the 1990-1994 period showing a frequency of 2.93 words per million and the 2015-2019 period showing a frequency of 8.93 words per million, which is approximately three times higher than the first period. Regarding the sources, the term is most frequently found in academic texts, probably due to its use in academic fields such as queer theory, queer studies, and queer linguistics.

SECTION	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
FREQ	5057	927	612	630	150	594	564	265	1315	355	381	416	563	707	1096
WORDS (M)	993	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8	121.1	125.2	124.6	123.1	123.3	122.8
PER MIL	5.09	7.21	4.93	4.92	1.19	5.02	4.47	2.18	10.98	2.93	3.04	3.34	4.58	5.73	8.93
SEE ALL SUB-SECTIONS AT ONCE															

Figure 1: Chart showing the frequency of the word “queer” in the COCA. On the left side the frequency is grouped by source, and on the right side it is grouped by five-year period.

To explore more in depth what types of texts exactly are contributing to these frequencies, we used the option to see all subsections at once on the website and copied the data into two new CSV files, one with information on the sources and one with information on the years. This data includes the detailed sources of the texts, which we sorted by words per million. The top 15 of this order can be found in Figure 2, where we can see that academic texts in humanities (ACAD:Humanities) are the source with the highest frequency of the word “queer”, more than twice the frequency of the next source, which is miscellaneous academic texts (ACAD:Misc).

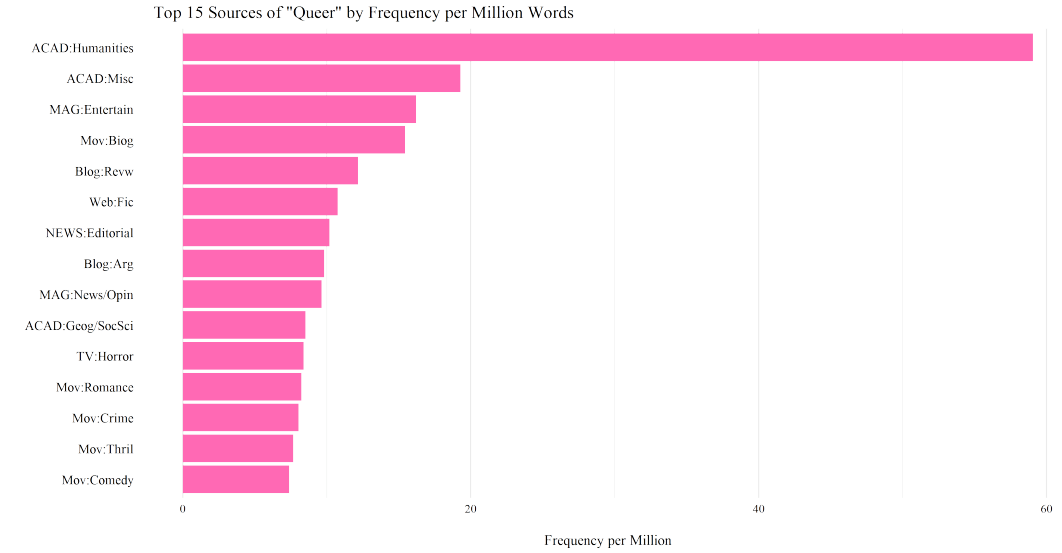

Figure 2: Chart of the frequency of the word “queer” in words per million (x axis) in the 15 sources (y axis) with the highest frequency.

Then, we explored the increase in the usage of “queer” by plotting the frequency in words per million for each five-year period and for each year, as seen in Figure 3. Even without the smoothing of the data that five-year groups bring, the increase in usage is clear.

However, these two lists of subsections do not include any information on the relation between sources and years, so to explore that we had to use the text dataset described in subsection 2.1. We plotted the frequency of the word “queer” for each five-year period and for each year as seen in Figure 4, with different colours representing the different sources of the texts where the word appears. Once again, we see that most of the usage and its increase comes from academic texts, as seen in Figure 2. However, regardless of the division into sources, there is a big difference in the shape of the bar charts now compared to those in Figure 3.

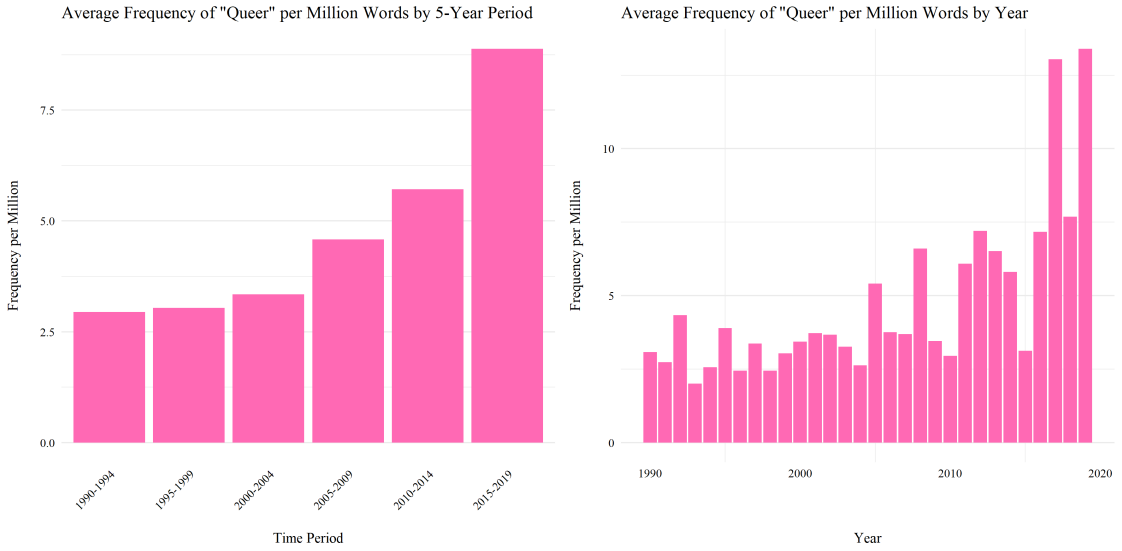

Figure 3: Charts of the frequency of the word "queer" in words per million (y axis) for each five-year period (x axis) on the left, and for each year (x axis) on the right.

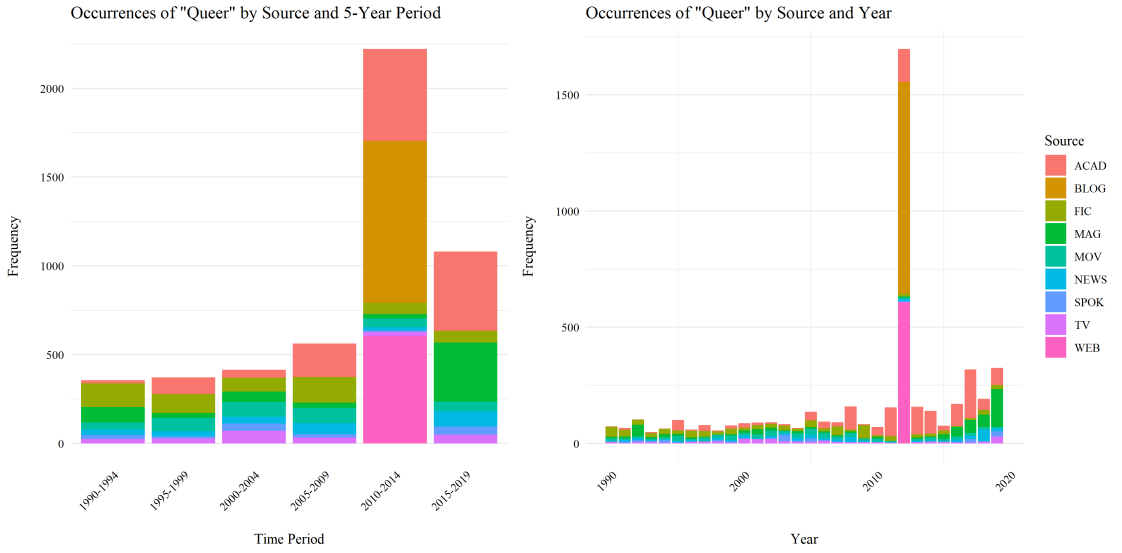

Figure 4: Charts of the frequency of the word "queer" (y axis) for each five-year period (x axis) on the left, and for each year (x axis) on the right, with colours representing different text sources.

We used the text dataset for these plots instead of the table of frequencies found under the chart option on the COCA website, and this difference implies that the table of frequencies may not correspond with the text dataset under the list option on the website. In the text dataset, all texts sourced from blogs and websites have been annotated as published in 2012. This is not consistent with the results plotted above and there is the chance that the dating on these texts is all wrong. Since data on dates is especially important to properly answer our research question and it is impossible to know if the texts sourced from blogs and websites have been correctly dated, we decided to exclude these texts from the dataset used for sentiment analysis.

2.3 Sentiment Analysis

All the texts sourced from blogs and websites were removed from the dataset before performing any sentiment analysis for the reasons explained in subsection 2.2 using the `filter()` function from the `dplyr` package (Wickham et al., 2023). To apply sentiment analysis to the resulting texts in our dataset we applied the `sentiment_by()` function from the `sentimentr` package (Rinker, 2021), as it has proved to return good results in previous studies (Mi & Zhan, 2023). This function requires sentences, so the texts were first pre-processed into lowercase using `tolower()` from base R and

punctuation was removed with the `removePunctuation()` function from the `tm` package (Feinerer, Hornik, & Meyer, 2008). They were then segmented into sentences using the `get_sentences()` function, also from the `sentimentr` package (Rinker, 2021). The `sentiment_by` function uses the Syuzhet lexicon (Jockers, 2015) to calculate a sentiment score in the range [-1, 1], where -1 is the most negative and 1 is the most positive (Jockers, 2015). It returns a data structure with the ID of each text, the ID of each sentence, the word count of each sentence, and the sentiment score of the text in the range [-1, 1]. We then selected only this last column, `ave_sentiment`, as it comprises the sentiment scores of all the texts in our dataset, and we saved it as a new column `sentiment`. The code used for this step can be found in Figure 5.

```
dfqueer_sentiment <- dfqueer %>%
  filter(!source %in% c("BLOG", "WEB")) %>%
  mutate(text = tolower(text)) %>%
  mutate(text = removePunctuation(text)) %>%
  mutate(sentences = get_sentences(text)) %>%
  mutate(sentiment = sentiment_by(sentences)$ave_sentiment) %>%
  select(-sentences)
```

Figure 5: Chunk of R code used to obtain sentiment scores for each text in the dataset through the `sentimentr` package.

To study how the sentiment associated with "queer" has changed over time, we calculated the average sentiment score across different time intervals, first for each five-year period and then for each year. We grouped the data by the relevant time variable, namely `period` or `year`, calculated the average sentiment `avg_sentiment` from the sentiment scores, and saved this into a new dataframe. This was done using the `group_by` and `summarise()` functions from the `dplyr` package and the `mean()` function from base R. The code used for this process is included in Figure 6 and the results can be found in Figure 10 in section 3.

```
df_avgsentperiod <- dfqueer_sentiment %>%
  group_by(period) %>%
  summarise(avg_sentiment = mean(sentiment, na.rm = TRUE), .groups = "drop")

df_avgsentyear <- dfqueer_sentiment %>%
  group_by(year) %>%
  summarise(avg_sentiment = mean(sentiment, na.rm = TRUE), .groups = "drop")
```

Figure 6: Chunk of R code used to obtain the average sentiment score for each five-year period and for each year.

The average sentiment scores represent a general tendency of the sentiments associated with "queer", and it depends on both the number and the intensity of the scores. However, it does not give us insights on how these scores are distributed across the data. For example, a balanced set of texts with some mildly positive and some mildly negative scores may result in the same average sentiment as another set of texts with many lightly positive and few strongly negative scores. A deeper analysis into the distribution and intensity of positive and negative sentiment scores is necessary to reach meaningful conclusions. Thus, we classified the texts into positive and negative according to the polarity of their sentiment score by creating a `sentiment_group` variable as seen in Figure 7, and we explored the distribution and sentiment intensity of these groups over time.

```
dfqueer_sent <- dfqueer_sent %>%
  mutate(sentiment_group = ifelse(sentiment >= 0, "Positive", "Negative"))
```

Figure 7: Chunk of R code used to create a `sentiment_group` variable that separates the dataset into Positive and Negative groups.

Note that the positive sentiment group is defined in the code in Figure 7 as having a sentiment score larger than or equal to 0, which means that the 16 texts with a sentiment score of exactly 0 are classified as positive. We chose not to create a separate sentiment group for the instances

where "queer" is used in a neutral context, as these only account for 0.459% of the dataset used for sentiment analysis.

After the texts were classified into positive and negative sentiment, we calculated the percentage of texts in each class across time intervals. We grouped the data by either `period` or `year` again, counted the number of texts in each sentiment group, calculated their percentage within each time interval, and saved the results in a new dataframe. This was done using the `group_by`, `summarise`, and `mutate()` functions from the `dplyr` package, along with base R functions. The code used for this process is included in Figure 8 and the results are presented in Figure 11 in section 3.

```
df_posnegpercentageperiod <- dfqueer_sentiment %>%
  group_by(period, sentiment_group) %>%
  summarise(count = n(), .groups = "drop_last") %>%
  mutate(percentage = count / sum(count))

df_posnegpercentageyear <- dfqueer_sentiment %>%
  group_by(year, sentiment_group) %>%
  summarise(count = n(), .groups = "drop_last") %>%
  mutate(percentage = (count / sum(count)) )
```

Figure 8: Chunk of R code used to calculate the percentage of positive and negative texts by time interval, first by five-year period and then by year.

Finally, we defined the sentiment intensity of a text as the absolute value of the sentiment score of the text and calculated the average sentiment intensity in each sentiment group across time intervals. We grouped the data by either `period` or `year`, then calculated the average intensity for each sentiment group, and saved the results in a new dataframe. This was done using the `mutate`, `group_by`, and `summarise` functions from the `dplyr` package once again. The code used for this process is shown in Figure 9, and the results are presented in Figure 12 in section 3.

```
df_posnegabsperiod <- dfqueer_sentiment %>%
  mutate(sentiment_abs = abs(sentiment)) %>%
  group_by(period, sentiment_group) %>%
  summarise(avg_sentiment = mean(sentiment_abs, na.rm = TRUE), .groups = "drop")

df_posnegabsyear <- dfqueer_sentiment %>%
  mutate(sentiment_abs = abs(sentiment)) %>%
  group_by(year, sentiment_group) %>%
  summarise(avg_sentiment = mean(sentiment_abs, na.rm = TRUE), .groups = "drop")
```

Figure 9: Chunk of R code used to calculate the intensity of positive and negative texts by time interval, first by five-year period and then by year.

3 Results

The average sentiment score across time are presented in Figure 10 below, where we can see that the sentiment associated with the word "queer" has become more positive throughout the thirty years that are covered in our dataset. If we look at the average score per five-year period, the results seem to be all positive and sentiment seems to mostly increase every period, except for the 2005-2008 period, which shows a significant decrease; and the 2015-2019 period, which shows a slight decrease in sentiment. However, if we look at the score per year, we can see that this increase is very much not homogeneous and that there are several years with a negative average sentiment score, namely 1992, 1995, 1999, 2001, and 2008.

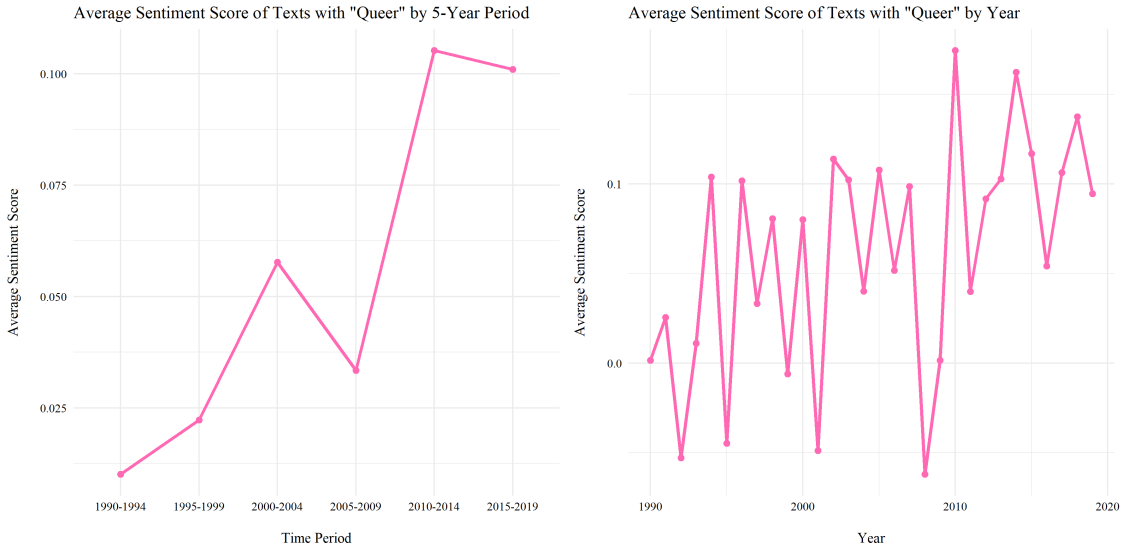

Figure 10: Average sentiment score (y axis) of texts in COCA containing the word "queer" for each five-year period (x axis) on the left, and for each year (x axis) on the right.

The percentage of texts with a positive and negative sentiment score across time can be found in Figure 11. Note that the two lines in these plots are always horizontally symmetrical, as there are only two sentiment groups and their percentages will always add up to 100%. Again, if we look only at the percentage per five-year period, it seems that positive texts always outnumber negative ones, and that the percentage of positive texts is constantly growing, except in the 2005-2009 and 2015-2019 periods, where it slightly decreases. If we look at the plot by years, we can see that the increase in percentage of positive texts is not that smooth and that there are several years where negative texts outnumber positive ones, namely 1992, 1999, 2001, and 2008.

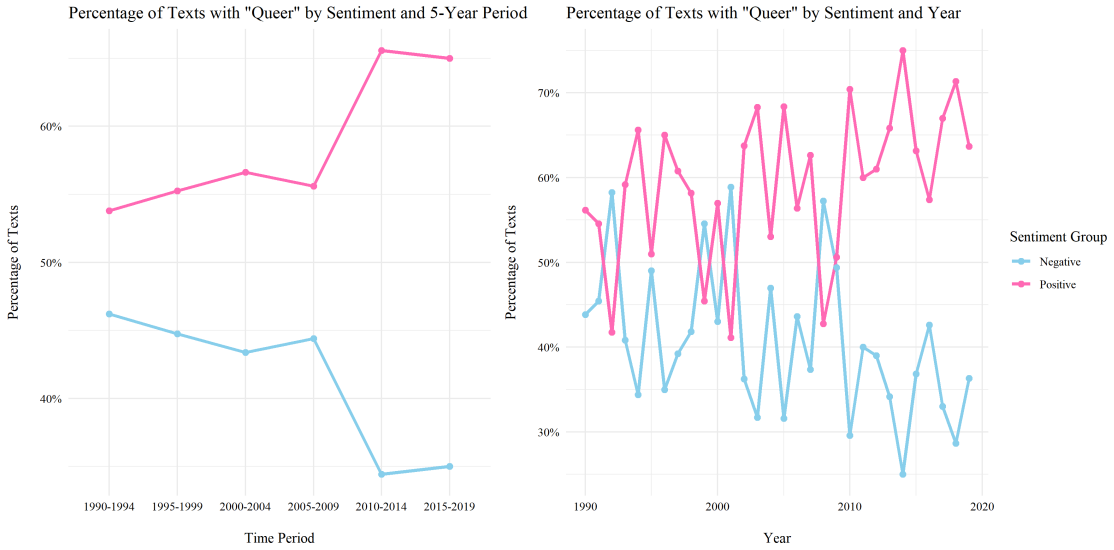

Figure 11: Percentage of texts (y axis) containing the word "queer" in the COCA with a positive (pink) and negative (blue) sentiment score for each five-year period (x axis) on the left, and for each year (x axis) on the right.

Finally, the average sentiment intensity of positive and negative scores across time is presented in Figure 12. In the first period of our dataset, 1990-1994, Negative texts published in 1990-1994, the first period of our dataset, are almost twice as intense as positive texts from the same period. In the following five-year period, the intensity of negative texts decreases, but they remain slightly more intense than positive ones. After that, the intensity of both positive and negative texts grows, but the intensity of positive texts remains greater than that of negative texts for the rest of the dataset. However, if we look at the plot by years, we see that in 1990, 1992, 1993, 1995, 1997, 2001,

2003, 2008, 2009, and 2011 intensity was higher in negative texts than in positive ones. The years with a higher intensity for both positive and negative texts were 2002, 2003, and 2018, meaning that the sentiment was more polarised.

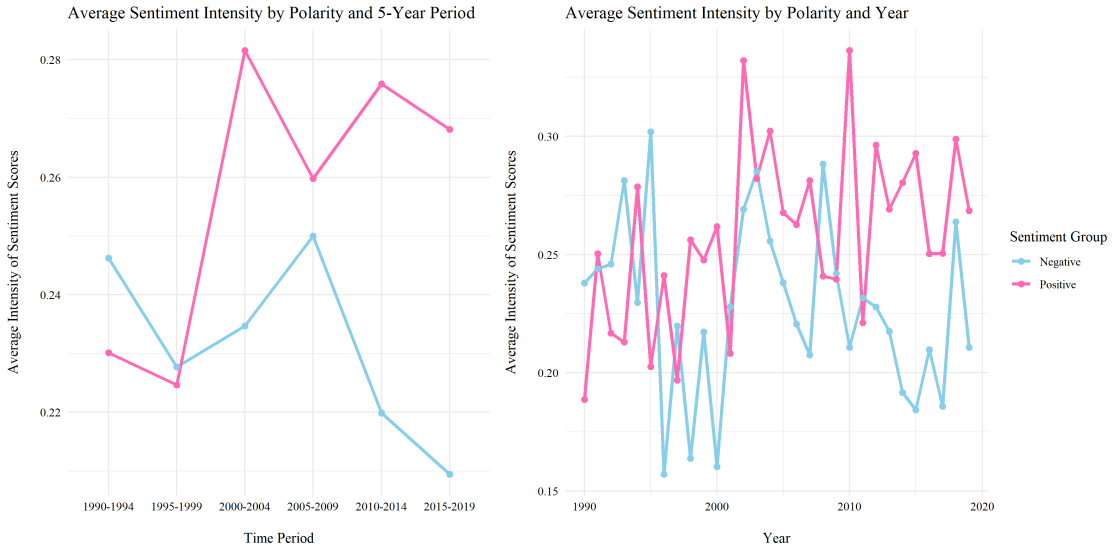

Figure 12: Average intensity of scores of texts in absolute value (y axis) containing the word "queer" in the COCA with a positive (pink) and negative (blue) sentiment score for each five-year period (x axis) on the left, and for each year (x axis) on the right.

The results on percentage and intensity of positive and negative texts across time are somewhat similar to the findings on the average sentiment score across time, as the latter is determined by both the amount and the value of the sentiment scores in each sentiment group.

4 Discussion

We expected the overall sentiment associated with the word "queer" to have significantly increased throughout the thirty years that the COCA encompasses, and our results align with this expectation. However, this increase in positive sentiment was not stable. There were several years where the average sentiment score was negative, because the negative texts in the corpus outnumbered the positive ones, or because the negative ones were more intense, or both. The most notable case is that of 2008, which shows the third highest percentage and the second highest intensity for negative texts, resulting in the lowest average sentiment score. Something similar happens in 2001, which shows the highest percentage of negative texts, so although not so intense, it ranks as the third lowest average sentiment scores. The other years showing most negative average sentiment scores are in the 1990's, which was expected, but the cases of 2008 and 2001 were surprising to find. We explored the evolution of the sentiment associated with "queer" over time, but the reasons behind these results are outside of the scope of this study. Thus, investigating the causes of this increased negative sentiment for "queer" in 2001 and 2008 and maybe even a possible link between queerphobia or conservativism and economical crises would be an interesting direction for future research.

Another limitation of this study is the lack of distinction between intra-group and inter-group use of the word "queer". Most studies on the question of "queer" focused on its use by people that identified as such (Gamson, 1995; Jacobs, 1998; Zosky & Alberts, 2016), and previous corpus-based analyses included information about whether the author identified as queer or not (Hanneder, 2023; Engra Minaya, 2024), which is not available in the COCA. Future research on the use and perception of "queer" and other slurs should include this information in their data, which means that the COCA would not be a good enough source of data on its own.

5 References

Cambridge University Press. (n.d.). Queer. In Cambridge Dictionary. Retrieved April 10, 2025, from `https://dictionary.cambridge.org/dictionary/english/queer`.

Rationale: We used the dictionary definition of the word "queer".

Davies, Mark. (2008-). The Corpus of Contemporary American English (COCA). Available online at `https://www.english-corpora.org/coca/`.

Rationale: We used this corpus as the dataset for our experiments.

Engra Minaya, S. (2024). Identidades sociolingüísticas y reapropiación: Análisis sociocultural de maricón y bollera en un corpus de Twitter. MariCorners: Revista de Estudios Interdisciplinares LGTBIA+ y queer, 1(1), 235–266. DOI.

Rationale: We used this paper for inspiration on corpus-based approaches to the study of the reclamation of "queer", although their methodology is different, but it is useful for future work.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. Journal of Statistical Software, 25(5), 1–54. DOI

Rationale: We used the `tm` package for text mining in R for the pre-processing of the data.

Gamson, J. (1995). Must Identity Movements Self-Destruct? A Queer Dilemma. Social Problems, 42(3), 390–407. DOI.

Rationale: We used this paper to learn more about the debate over the use of the term "queer" in the 1990's, as that is the starting point of our study.

Hanneder, H., & Best, S. (2023). Did dykes die out or where have they gone?: An interdisciplinary analysis of a key term in queer archives. Tijdschrift Voor Genderstudies, 26(3/4), 313–334. DOI.

Rationale: We used this paper for inspiration when formulating our research question and methodology, as the author also uses a corpus-based approach with COCA to study the reclamation of a slur through sentiment analysis among other methods, although in this case it is the term "dyke" and their findings are quite different from ours. In fact, we were more interested in "dyke" than in "queer" for personal reasons, but their analysis goes beyond the scope of ours and the humbling nature of their results made us opt for a more widely reclaimed term such as "queer" for our project.

Jacobs, G. (1998). The struggle over naming: A case study of 'queer' in Toronto, 1990–1994. World Englishes, 17(2), 193–201. DOI.

Rationale: We used this paper to learn about the use and perception of the word "queer" in the 1990's, as that is the time period for the oldest texts in the COCA and thus the starting point for our study on the change of the sentiment associated with the word.

Jockers, M. L. (2015). Syuzhet: Extract sentiment and plot arcs from text. Github.

Rationale: The Syuzhet lexicon was used to obtain the sentiment scores through the package `sentimentr`.

Merriam-Webster. (n.d.). Queer. In Merriam-Webster.com dictionary. Retrieved April 10, 2025, from `https://www.merriam-webster.com/dictionary/queer`.

Rationale: We took the usage note from the dictionary to explain how "queer" has been and is used and how dictionaries record this use.

Mi, Z., & Zhan, H. (2023). Text Mining Attitudes toward Climate Change: Emotion and Sentiment Analysis of the Twitter Corpus. DOI.

Rationale: We used this paper as inspiration for our methodology, as they also performed sentiment analysis to explore the perception of a sociopolitical issue, namely climate change, and obtained good results.

Motschenbacher, H. (2022). Linguistic Dimensions of Sexual Normativity: Corpus-Based Evidence. Routledge. DOI.

Rationale: We used this book for inspiration on the process of choosing a corpus and a methodology. More specifically, we used the case study in section 6.1. where the author uses the COCA corpus to perform a co-occurrence analysis on "queer" and other related words.

Rinker, T. W. (2021). sentimentr: Calculate text polarity sentiment (Version 2.9.1). Buffalo, New York. Github.

Rationale: We used the `sentimentr` package to obtain the sentiment scores used in our analyses.

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). dplyr: A grammar of data manipulation (R package version 1.1.4). https://dplyr.tidyverse.org

Rationale: We used the `dplyr` package in R several times to process the data.

Wilkinson, M. (2022). Radical contingency, radical historicity and the spread of 'homosexualism': A diachronic corpus-based critical discourse analysis of queer representation in The Times between 1957–1967 and 1979–1990. Discourse, Context & Media, 48, 100623. DOI.

Rationale: We used this paper as inspiration for our research question and methodology, as they also use a corpus-based approach to study the perception of gayness and the gay community, although they use collocation analysis instead of sentiment analysis and focus on the association of words like "homosexual" and "AIDS".

Zosky, D. L., & Alberts, R. (2016). What's in a name? Exploring use of the word queer as a term of identification within the college-aged LGBT community. Journal of Human Behavior in the Social Environment, 26(7–8), 597–607. DOI.

Rationale: I used this paper to learn about the use and perception of the word "queer" in 2016. Although the most recent texts in the COCA are from 2019, this is the latest paper I found that explores the emotional reaction to the word "queer" by young American English speakers.