Tracking Sanitation in Historic American Cookbooks

Adrian Lee

2025-12-10

Table of contents

Abstract

This project examines how the language of sanitation changed in American cookbooks between 1800-1920. Using the Feeding America: Historic American Cookbook Dataset (74 cookbooks),

I focus on a small lexicon of "general" cleanliness terms (e.g., boil, wash, clean) and scientific sanitation terms (e.g., sterilize, disinfect). After tokenizing and normalizing the corpus, I analyze term distributions, track frequencies over 4 chronological periods, compute log-likelihood keyness for early vs. later cookbooks, and compare collocations for clean and sterilize. The results show that general cleanliness vocabulary is pervasive throughout the period, while scientific terms are rare and only begin to appear in the later cookbooks, where sterilize gradually increases (but never approaches the frequency of older terms.) Keyness and collocation analyses suggest that later texts emphasize boiling and washing as distinct hygienic procedures in a technical, equipment-focused manner (i.e. jars), whereas clean is embedded in a broader housekeeping network. Overall, the study indicates that germ-theory sanitation practices entered cookbook language as a minor addition on top of long-standing domestic cleanliness discourse, supplementing rather than replacing it.

1 Introduction

This project investigates how the language of sanitation and food safety has evolved in American cookbooks between 1800 and 1920. Using the corpus, Feeding America: Historic American Cookbook Collection, I examine how instructions related to cleanliness, preservation, and hygiene (e.g., "wash," "boil," "sterilize," "preserve") change over time through these hundreds of digitized cookbooks. I am especially interested in when and how more explicitly scientific sanitation terms, such as sterilize, disinfect, and sanitize, enter the culinary vocabulary and potentially replace earlier, more general terms like clean, pure, or fresh (or perhaps was never mentioned either.) By tracing these lexical shifts, I aim to see not only whether sanitation-related language becomes more frequent, explicit, and complex, but also how cookbooks frame cleanliness as a moral, domestic, or scientific duty over time. Ultimately, this study treats cookbooks as both practical guides and cultural artifacts that mirror societal developments in public health, germ theory, and modern food safety practices.

Furthermore, I look to build on the existing work of historians of food and domestic science. Specifically, I want to focus on how 19th-20th century cookbooks were central tools for teaching individuals emerging ideals of cleanliness and order, especially as germ theory reshaped public understandings of disease and contamination. Scholars have also argued that these texts helped construct gendered norms of "good housekeeping," casting cleanliness as both a moral and civic duty. Finally, corpus-based studies of historical English have demonstrated how shifts in everyday instructional genres can make broader scientific and cultural change visible at the level of lexis and collocation. By bringing these perspectives together and applying quantitative text analysis to the cookbook corpus, this project contributes a more systematic, language-focused view of how public health ideas entered the kitchen through printed recipes and household advice.

2 Data

The data for this project comes from the course-given corpus, Feeding America: The Historic American Cookbook Dataset, which consists of transcribed and encoded text from 76 American cookbooks held in the Michigan State University Libraries' Special Collections. These cookbooks were selected from a much larger collection of more than 7,000 volumes as representative of key periods and themes in American cookbook history, spanning roughly the late 18th through the early 20th century. The full dataset includes plain-text transcriptions as well as XML files for recipes, recipe types, ingredients, and cooking implements, slightly cleaned to remove tables of contents and other front matter.

Furthermore, I work with the plain-text portion of the dataset (74 cookbook text files), treating each cookbook as a single document and focusing on those published between about 1800 and 1920, the period in which germ theory and modern food safety practices emerge. At the same time, the corpus reflects published, often middle-class, perspectives and may over represent certain regions or authors, so any findings about "American" cooking and cleanliness should be read as patterns in influential print cookbooks rather than in all domestic practice.

Table 1: Overview of the Feeding America cookbook corpus used in this study.

corpus	Files	Words (tokens)
cookbooks	74	4200176

3 Methods

All analyses were conducted in R using the quanteda package. Because this study focuses on sanitation-related vocabulary, I first compiled a targeted lexicon of terms associated with cleanliness and food safety. This list has common terminology (wash, boil, clean, fresh, pure) with more explicitly scientific terms linked to germ theory and modern preservation practices (sterilize, disinfect, sanitize). Each of the 74 cookbooks is treated as a single document. I tokenized the texts, removed punctuation, numbers, and English stopwords, and then calculated token counts for the sanitation lexicon in each document. To approximate historical time, I used the publication years available in the dataset to divide the corpus into four chronological periods from the earliest to the latest cookbooks. Within each period, I computed normalized frequencies (per 10,000 words) for each sanitation term so that differences in vocabulary are not driven by differences in corpus size.

I then combined these steps with three complementary analytical methods. First, I used a time-series analysis: for each period, I plot the normalized frequencies of general terms (wash, boil, clean, fresh, pure) and scientific terms (sterilize, disinfect, sanitize). This allows me to trace when explicitly scientific sanitation vocabulary first appears and whether it increases over time, addressing my questions about lexical emergence and possible replacement of earlier cleanliness terms. Second, I use log-likelihood keyness to compare earlier versus later subcorpora. Using the sanitation lexicon as the focus set, I calculate G^2 statistics that identify which

sanitation-related words are statistically more distinctive in later cookbooks than in earlier ones. This comparison highlights which practices (e.g., boiling, washing, sterilizing) become especially characteristic of later texts rather than simply more frequent overall.

Lastly, I conducted collocation analyses for selected terms such as clean and sterilize in early and later periods, using surrounding words to identify their most frequent lexical partners. This showed how the lexical environments of sanitation terms shifted from general housekeeping frames (e.g., sweeping, scrubbing) toward more technical, equipment-oriented or preservation-oriented language (e.g., jars, bottles, utensils), which speaks directly to my question about changing moral, domestic, and scientific framings of cleanliness. Throughout, I interpret patterns cautiously for very low-frequency scientific terms (especially sanitize) and acknowledge that the broad periodization can only approximate the timing of change rather than pinpoint exact years.

4 Results

4.1 Distributions

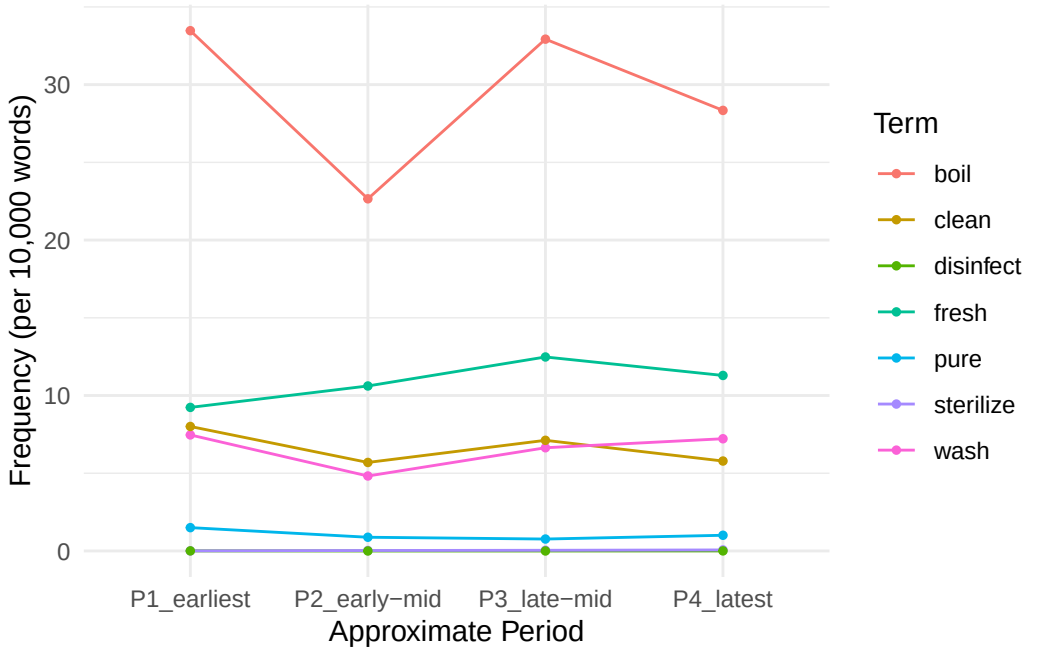
Table 2: Distribution of sanitation-related terms in the Feeding America cookbook corpus.

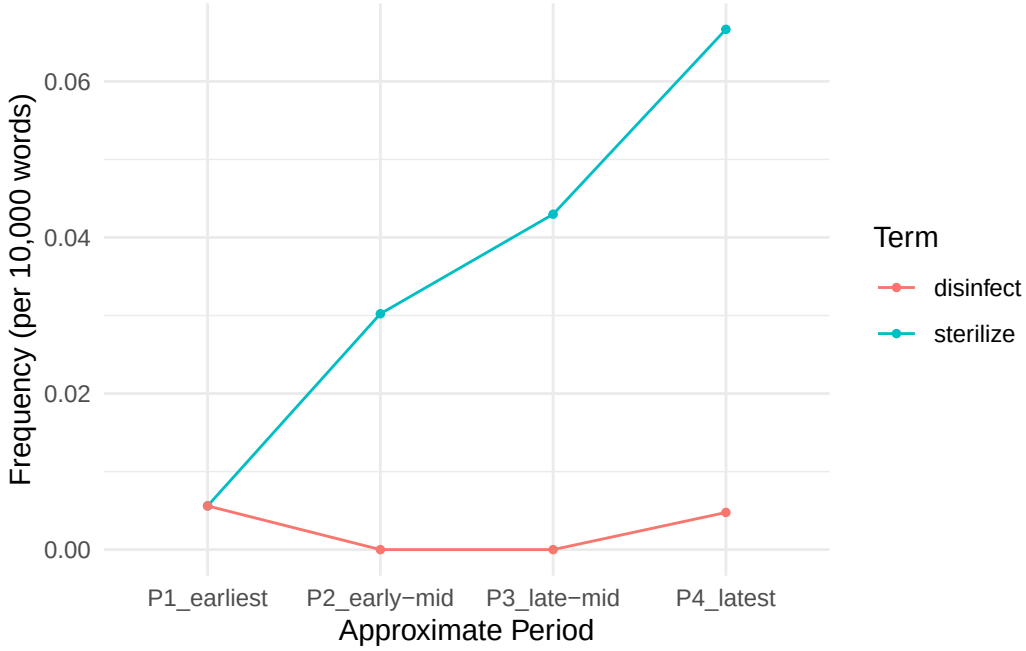
term	raw_count	category	per_10000
boil	21805	general	51.9144912
fresh	8099	general	19.2825253
clean	4912	general	11.6947480
wash	4883	general	11.6257033
pure	769	general	1.8308757
sterilize	28	scientific	0.0666639
disinfect	2	scientific	0.0047617

Table 3: Overall frequency of general vs scientific sanitation terms.

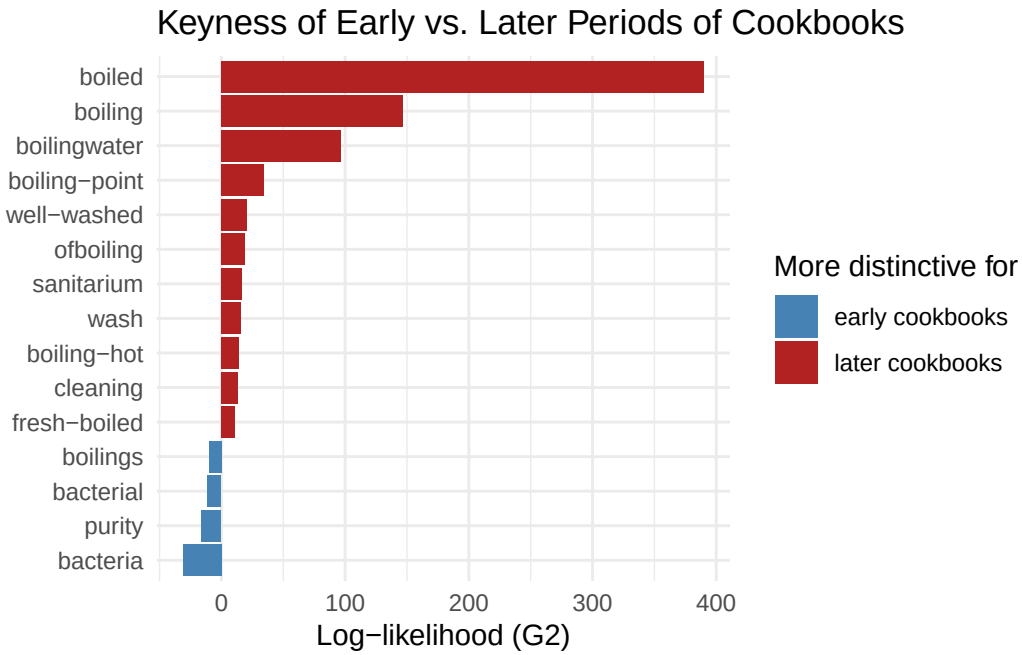
category	total_count	per_10000
general	40468	96.3483435
scientific	30	0.0714256

4.2 Time-Series

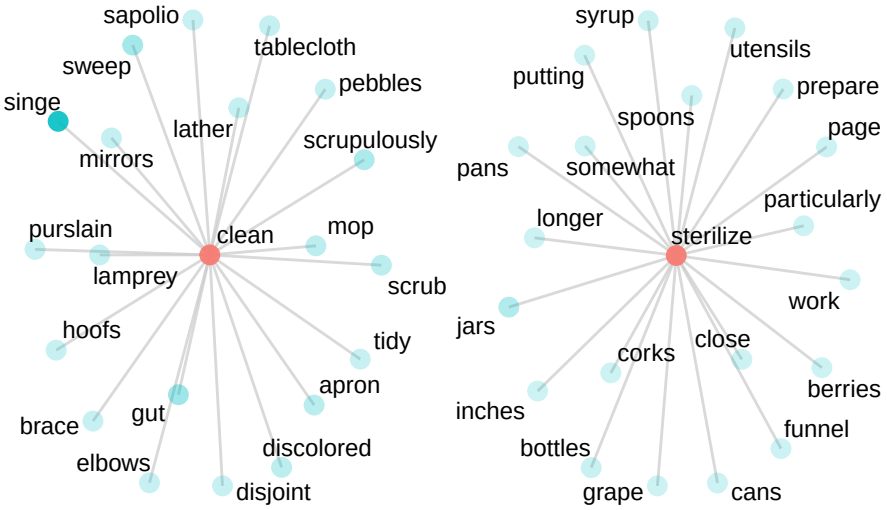




4.3 Log-Likelihood Keyness



4.4 Collocation Analyses



5 Discussion

5.1 Main Findings

The results suggest that explicitly scientific sanitation terms enter American cookbook language only quite late, and that they supplement rather than entirely replace the cleanliness vocabulary bank. The distributional results in Tables 2-3 show that general sanitation terms dominate the corpus, with boil alone occurring over 50 times per 10,000 words and fresh, clean, and wash all in the 10-20 (per-10,000) range. In contrast, the scientific terms sterilize and disinfect appear only 30 times in total across more than four million tokens which is roughly 0.07 instances per 10,000 words. This imbalance suggests that, even by 1920, cookbook writers relied primarily on traditional cleanliness vocabulary rather than the newer technical lexicon.

Next, the time-series plots build on this idea by addressing when scientific language begins to matter. General terms such as boil, fresh, clean, and wash are present in every period and remain relatively stable, with boil consistently one of the most frequent sanitation verbs (Figure 4.2). In contrast, sterilize is absent for the majority of the earliest period and then increases consistently across the later three periods, while disinfect appears only sporadically at low levels (Figure 4.3). These findings support the idea that technical sanitation vocabulary is a late addition tied to the era of germ theory: it enters the cookbook corpus in small instances rather than transforming the lexicon suddenly and widespread.

Then, keyness and collocation analyses help explain how these new terms reshape the discourse of cleanliness. The keyness results for early versus later cookbooks highlight a cluster of boiling-related forms, such as boiled, boiling, boilingwater, boiling-point, etc, and phrases, such as well-washed and cleaning, as especially distinctive of later texts, while terms like bacteria and purity are more characteristic of earlier cookbooks (Figure 4.3). This suggests that later authors do not only mention cleanliness more often; they actually foreground specific procedures, especially boiling and thorough washing, as important practices. The collocation networks reemphasizes this: in earlier texts, clean sits in a dense housekeeping network that includes mop, sweep, scrub, tidy, tablecloth, mirrors, and body-related terms like elbows and gut (Figure 4.4, left). This is most likely attributed to how cleanliness is a broad domestic action applied to floors, fabrics, bodies, and surfaces. In contrast, sterilize in later cookbooks is surrounded by jars, bottles, cans, corks, utensils, funnel, inches, and syrup (Figure 4.4, right), placing sanitation in the space of canning and equipment preparation. In other words, the newer scientific vocabulary appears primarily in specific contexts where readers are instructed to treat containers and tools in precise, almost laboratory-like ways.

Therefore, these findings have several implications for how we understand cookbooks as public-health texts. First, they suggest that 19th-20th century cookbooks did not abruptly switch from moral/domestic to scientific framings of cleanliness. Instead, they supplemented germ-theory practices onto an already existing discourse of household order and cleanliness. The general terms still remain the foundation of sanitation talk, but in later texts they are joined by a small set of technical verbs and by more detailed descriptions of boiling, washing, and equipment treatment. Second, the fact that sanitize never meaningfully appears, and disinfect barely does, hints that public-health vocabulary may have circulated more in medical

or governmental texts than in domestic instructional writing. For home cooks, boiling and sterilizing jars or other equipment may have been the most important linguistic idea for germ-theory ideas.

5.2 Limitations & Next Steps

However, I must acknowledge some limitations regarding these conclusions. First, the corpus utilized consists of only 74 cookbooks, and while they are historically significant, they represent published, often middle-class perspectives rather than everyday kitchen practice. My periodization used is approximate and based on grouping rather than exact dates, which restricts how precisely I can find the "arrival" of particular terms. More importantly, the scientific sanitation lexicon showed low-frequency: patterns in sterilize and disinfect are suggestive but small, and the absence of the word sanitize may reflect corpus size, genre, or regional selection rather than a complete lack of usage. In the future, I perhaps hold to expand on my built lexicon and incorporate parts-of-speech tagging to help capture a wider variety of these terms.

Nevertheless, despite these constraints, the analyses provide a useful proof of concept for treating cookbooks as windows into linguistic and scientific change. Future work could better the ambiguity surrounding time periods by finding precise publication years or by comparing this corpus with medical pamphlets, housekeeping manuals, or regional cookbooks to see whether scientific sanitation language appeared earlier in other places. Furthermore, a geospatial analysis of publication places could be interesting if possible; it might reveal whether urban places adopted technical sanitation vocabulary earlier than rural regions. Overall, this study suggests that germ-theory ideas seeped into American kitchen discourse unevenly: not by sweeping away older cleanliness terms, but by reconfiguring what it meant to boil, wash, and "sterilize" the tools and containers of domestic food production.

6 Acknowledgments

I utilized Generative AI to assist me in understanding the different usages of each methodology (keyness, collocation, time-series, stylometric classifiers, etc) from lectures and classwork to help in my decision-making of which methods to choose. Furthermore, I also used it to revise specific errors in my graphs, specifically sanitaiton-related keyness which found the keyness for the general corpus and not for sanitation-related terms. Lastly, I would state that using Generative AI was overall helpful in constructing and choosing which methods to use for my report, providing a high-level overview of topics that would have taken longer to look at.

7 Works Cited

Michigan State University Libraries, Stephen O. Murray & Keelung Hong Special Collections.(n.d.). Feeding America: The Historic American Cookbook Dataset. https://lib.msu.edu/feedingamericadata/

Brezina, Vaclav. Statistics in Corpus Linguistics: A Practical Guide. Cambridge: Cambridge University Press, 2018.

8 Code (for Reference)

```
#Distribution Table
library(tidyverse)
library(quanteda)
library(knitr)
library(kableExtra)

#Sanitation lexicon
general_terms <- c("wash", "boil", "clean", "fresh", "pure")
scientific_terms <- c("sterilize", "disinfect", "sanitize")
sanitation_terms <- c(general_terms, scientific_terms)

san_dfm <- dfm_keep(dfmat, pattern = sanitation_terms)
total_tokens<- sum(ntoken(dfmat))

dist_overall <- tibble(
  term= colnames(san_dfm),
  raw_count = colSums(san_dfm)) %>%
  mutate(
    category   = case_when(
      term %in% general_terms ~ "general",
      term %in% scientific_terms ~ "scientific",
      TRUE ~ NA_character_),
    per_10000 = (raw_count / total_tokens) * 10000) %>%
  arrange(desc(raw_count))

kable(
  dist_overall,
  format  = "latex",
  booktabs = TRUE,
  caption = "Distribution of sanitation-related terms in the Feeding America cookbook corpus
  kable_styling(latex_options = "HOLD_position", font_size = 9)
```

```
dist_category <- dist_overall %>%
  group_by(category) %>%
  summarise(
    total_count = sum(raw_count),
```

```
    per_10000 = (total_count / total_tokens) * 10000,
    .groups= "drop")

kable(
  dist_category,
  format = "latex",
  booktabs = TRUE,
  caption = "Overall frequency of general vs scientific sanitation terms.") %>%
  kable_styling(latex_options = "HOLD_position", font_size = 9)
```

```
#Time-Series
library(readtext)
library(quanteda)
library(dplyr)
library(tidyr)
library(ggplot2)
cookbooks <- readtext("cookbook_corpus/*.txt") %>%
  arrange(doc_id) %>%
  mutate(doc_index = row_number())

cb_corpus <- corpus(cookbooks)
cb_tokens <- tokens(
  cb_corpus,
  include_docvars = TRUE,
  remove_punct= TRUE,
  remove_numbers= TRUE,
  remove_symbols  = TRUE,
  what = "word"
) |>
  tokens_tolower()

dfmat <- dfm(cb_tokens)

#Defining the 4 periods
n_periods <- 4
docvars(dfmat, "period") <- cut(
  docvars(dfmat, "doc_index"),
  breaks = quantile(docvars(dfmat, "doc_index"),
                    probs = seq(0, 1, length.out = n_periods + 1)),
  include.lowest = TRUE,
  labels = c("P1_earliest", "P2_early-mid", "P3_late-mid", "P4_latest"))
```

```r
sanitation_lex <- c(
  "wash", "boil", "clean", "pure", "fresh",
  "sterilize", "disinfect", "sanitize") #lexicon again

#Grouping by Period
dfm_period_all <- dfm_group(dfmat, groups = docvars(dfmat, "period"))
period_tokens <- rowSums(dfm_period_all)
san_terms_present <- sanitation_lex[sanitation_lex %in% featnames(dfm_period_all)]

dfm_period_san <- dfm_keep(
  dfm_period_all,
  pattern = san_terms_present,
  valuetype = "fixed")

period_stats <- as_tibble(as.matrix(dfm_period_san), rownames = "period") %>%
  pivot_longer(
    cols= -period,
    names_to = "term",
    values_to = "period_freq") %>%
  mutate(
    period_tokens = period_tokens[match(period, names(period_tokens))],
    freq_per_10k  = (period_freq / period_tokens)*10000)
```

```r
ggplot(period_stats,
  aes(x = period, y = freq_per_10k, color = term, group = term)) +
  geom_line() +
  geom_point(size = 1) +
  labs(main = "Frequency per Period",
    x = "Approximate Period",
    y = "Frequency (per 10,000 words)",
    color = "Term"
  ) +
  theme_minimal()
```

```
Ignoring unknown labels:
* main : "Frequency per Period"
```

```r
#"scientific" terms
scientific_terms <- c("sterilize", "disinfect", "sanitize")
period_stats %>%
  filter(term %in% scientific_terms) %>%
  ggplot(aes(x = period, y = freq_per_10k, color = term, group = term)) +
```

```
  geom_line() +
  geom_point(size = 1) +
  labs(
    main = "Frequency per Period (Scientific Terms)", #fix title later?
    x = "Approximate Period",
    y = "Frequency (per 10,000 words)",
    color = "Term") +
  theme_minimal()
```

```
Ignoring unknown labels:
* main : "Frequency per Period (Scientific Terms)"
```

```
#Log-Likelihood Keyness
library(readtext)
library(quanteda)
library(quanteda.textstats)
library(dplyr)
library(stringr)

cookbooks <- readtext("cookbook_corpus/*.txt") %>%
  arrange(doc_id) %>%
  mutate(
    doc_index = row_number(),
    period2   = ifelse(doc_index <= median(doc_index), "early", "later"))
cb_corpus <- corpus(cookbooks)

cb_tokens <- tokens(
  cb_corpus,
  include_docvars = TRUE,
  remove_punct = TRUE,
  remove_numbers = TRUE,
  remove_symbols  = TRUE,
  what = "word"
) |>
  tokens_tolower()

dfmat <- dfm(cb_tokens)
key_early_late <- textstat_keyness(
  dfmat,
  target  = docvars(dfmat, "period2") == "later",
  measure = "lr")
```

```r
library(ggplot2)
sanitation_pattern <- "(wash|washed|washing|
                        boil|boiled|boiling|
                        clean|cleaned|cleaning|
                        pure|purity|
                        fresh|freshness|
                        steriliz|disinfect|sanit|sanitary|sanitation|
                        germ|bacteria)"

key_sanitation <- key_early_late %>%
  filter(str_detect(feature, sanitation_pattern)) %>%
  arrange(desc(abs(G2)))

top_key_sanitation <- key_sanitation %>%
  slice_max(abs(G2), n = 15, with_ties = FALSE) %>%
  mutate(
    direction = ifelse(n_target > n_reference,
                       "later cookbooks",
                       "early cookbooks"))

ggplot(top_key_sanitation,
       aes(x = reorder(feature, G2), y = G2, fill = direction)) +
  geom_col(show.legend = TRUE) +
  coord_flip() +
  labs(title = "Keyness of Early vs. Later Periods of Cookbooks",
    x = NULL,
    y = "Log-likelihood (G2)",
    fill= "More distinctive for"
  ) +
  scale_fill_manual(values = c("early cookbooks" = "steelblue",
                               "later cookbooks" = "firebrick")) +
  theme_minimal()
```

```r
#Collocation Analyses
library(quanteda)
library(quanteda.extras)
library(dplyr)
library(stringr)
library(ggraph)

corpus_tokens <- cb_tokens %>%
  tokens_remove(stopwords("en"))
```

```r
tok_early <- tokens_subset(corpus_tokens, period2 == "early")
tok_later <- tokens_subset(corpus_tokens, period2 == "later")
coll_clean_early <- collocates_by_MI(tok_early, "clean")
coll_clean_later <- collocates_by_MI(tok_later, "clean")
coll_clean_early <- coll_clean_early %>%
  filter(col_freq >= 3) %>%
  arrange(desc(PMI)) %>%
  slice_head(n = 20)

coll_clean_later <- coll_clean_later %>%
  filter(col_freq >= 3) %>%
  arrange(desc(PMI)) %>%
  slice_head(n = 20)

coll_sterilize_later <- collocates_by_MI(tok_later, "sterilize") %>%
  filter(col_freq >= 2) %>%
  arrange(desc(PMI)) %>%
  slice_head(n = 20)

#Network plot
net_clean_sterilize <- col_network(coll_clean_early, coll_sterilize_later)
ggraph(net_clean_sterilize, layout = "stress") +
  geom_edge_link(color = "gray80", alpha = .75) +
  geom_node_point(aes(alpha = node_weight, color = n_intersects), size = 3) +
  geom_node_text(aes(label = label), repel = TRUE, size = 3) +
  scale_alpha(range = c(0.2, 0.9)) +
  theme_graph() +
  theme(legend.position = "none")
```