Brief introduction

So, lots of linguistic variation happening in real time. `coronavirus`, `covid19`, `pandemic`, and more recently (`the`?) `coronavirus pandemic`. For sure, these expressions are not proper synonyms – each refer to different "aspects" of the virus. `coronavirus` ~ virus. `covid19` ~ disease. `pandemic` ~ social/epi. Here, we take a super quick look at how this variation in reference is materializing on Twitter among the 535 voting members of the United States Congress since January 2020.

Twitter details

First things first, we obtain Twitter handles and some relevant biographical details (here, political affiliation) for the 100 US Senators and the 435 members of the House of Representatives from the unitedstates project.

```
library(tidyverse)
leg_dets <- 'https://theunitedstates.io/congress-legislators/legislators-current.csv'

twitters <- read.csv((url(leg_dets)),
                     stringsAsFactors = FALSE) %>%
  #filter(type == 'rep') %>% # & twitter!=''
  rename (state_abbrev = state,
          district_code = district)
```

Then we scrape the last 1000 tweets for each of the 535 members of congress using the `rtweet` package. Here, we are just trying to get all tweets from 2020 – 1,000 is overkill. We exclude re-tweets. The scraping process takes roughly an hour or so.

```
congress_tweets <- rtweet::get_timeline(
  twitters$twitter,
  n = 1000,
  check = FALSE) %>%
  mutate(created_at = as.Date(gsub(' .*$', '',
                                   created_at))) %>%
  filter(is_quote == 'FALSE' &
           is_retweet == 'FALSE' &
           created_at >= '2020-01-01' &
           display_text_width > 0)

# setwd("/home/jtimm/jt_work/GitHub/data_sets")
# saveRDS(congress_tweets, 'cong2020_tweets_tif.rds')
```

Then we join the two data sets. And calculate total tweets generated by members of Congress by party affiliation in 2020.

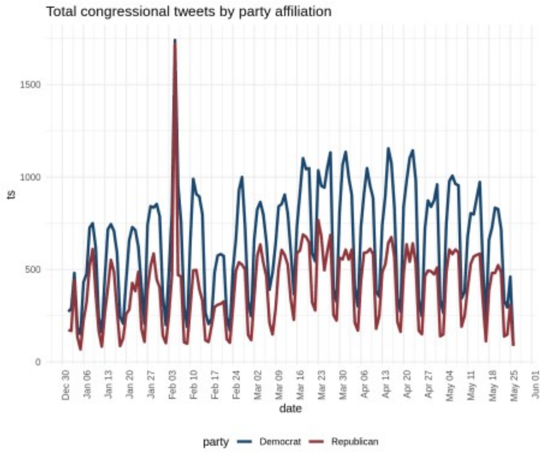
```
congress_tweets1 <- congress_tweets %>%
  mutate(twitter = toupper(screen_name)) %>%
  select(status_id, created_at, twitter, text) %>%
  inner_join(twitters %>% mutate(twitter = toupper(twitter)))


all_tweets <- congress_tweets1 %>%
  group_by(created_at, party) %>%
  summarise(ts = n()) %>%
  rename(date = created_at)
```

The figure below summarizes total tweets by party affiliation since the first of the year. Donald Trump presented his State of the Union address on February 5th, hence the spike in activity. There seems to be a slight upward trend in total tweets – perhaps one that is more prevalent among Democrats – presumably in response to the Coronavirus.

Also, Democrats do tweet more, but they also have numbers at present. And it seems that members of Congress put their phones down a bit on the weekends.

```
all_tweets %>%
  filter(party != 'Independent') %>% # Justin Amash & Bernie Sanders & Angus King
  ggplot() +
  geom_line(aes(x = date,
                y= ts,
                color = party
                ),
            size = 1.25) +
  theme_minimal() +
  ggthemes::scale_color_stata() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  scale_x_date(date_breaks = '1 week', date_labels = "%b %d") +
  theme(legend.position = 'bottom')  +
  labs(title = 'Total congressional tweets by party affiliation')
```



2019 NOVEL CORONAVIRUS & lexical variation

For clarity purposes, we will refer to the broad conceptual category of **the virus** in all caps as `2019 NOVEL CORONAVIRUS`. In contrast, the ways that speakers/tweeters can refer to this concept will be represented in lowercase, eg, `pandemic` & `covid19`. From this perspective, `coronavirus` is one way speakers can refer to the concept `2019 NOVEL CORONAVIRUS`.

So, with data sets in tow, the first step is to identify and extract the different lexical forms used to reference the virus. Referents include (1) `pandemic`, (2) `coronavirus`, `corona virus`, (3) `covid19`, `covid`, `covid 19`, `covid-19`, and (4) `coronavirus pandemic`. Some spelling variants. For simplicity, we ignore variation in orthographic case.

```
pan <- 'pandemic|'
cv <- 'coronavirus|corona virus|'
covid <- 'covid19|covid|covid 19|covid-19|'
cvp <- 'coronavirus pandemic'
searches <- paste0(pan, cv, covid, cvp)

covid_tweets <- lapply(1:nrow(congress_tweets1), function(x) {

    spots <- gregexpr(pattern = searches, congress_tweets1$text[x], ignore.case=TRUE)
    covid_gram <- regmatches(congress_tweets1$text[x], spots)[[1]]

    if (-1 %in% spots){} else {
      data.frame(doc_id = congress_tweets1$status_id[x],
                 date = congress_tweets1$created_at[x],
                 twitter = congress_tweets1$twitter[x],
                 party = congress_tweets1$party[x],
                 covid_gram = covid_gram,
                 stringsAsFactors = FALSE)}  })  %>%
  data.table:::rbindlist()
```

Attested variants are highlighted below. So, some disagreement on how things should be spelled. (It will be curious to see if this settles some moving forward, and conventions established.)

```
table(covid_tweets$covid_gram)
```

```
## 
##         corona virus         Corona virus         Corona Virus 
##                    4                    3                    3 
##         CORONA VIRUS          coronavirus          Coronavirus 
##                    1                 6757                 2493 
##          CoronaVirus          CORONAVIRUS coronavirus pandemic 
##                   94                   14                  340 
## Coronavirus pandemic Coronavirus Pandemic CoronaVirus Pandemic 
##                   45                    7                    1 
##                covid                Covid                COVID 
##                    7                  126                  674 
##             COVID 19             covid-19             Covid-19 
##                    6                    5                   11 
##             COVID-19              covid19              Covid19 
##                 1962                   48                   55 
##              COVID19             pandemic             Pandemic 
##                 3676                 1139                  111 
##             PANDEMIC 
##                    1
```

After normalizing spelling variation, a portion of the resulting table (less the tweet id) is presented below:

```
covid_tweets <- covid_tweets %>%
  mutate(covid_gram = tolower(covid_gram),
         covid_gram = ifelse(grepl('covid', covid_gram), 'covid19', covid_gram),
         covid_gram = ifelse(grepl('corona virus', covid_gram), 'coronavirus', covid_gram))

covid_tweets %>% sample_n(10) %>% select(-doc_id) %>%knitr::kable()
```

date	twitter	party	covid_gram
2020-03-20	REPBONAMICI	Democrat	coronavirus pandemic
2020-03-20	REPCARBAJAL	Democrat	covid19

date	twitter	party	covid_gram
2020-03-24	REPLINDASANCHEZ	Democrat	coronavirus
2020-03-06	REPLLOYDDOGGETT	Democrat	covid19
2020-03-27	REPPETEKING	Republican	pandemic
2020-03-16	LEADERHOYER	Democrat	coronavirus
2020-03-11	SENBRIANSCHATZ	Democrat	coronavirus
2020-03-20	CHELLIEPINGREE	Democrat	covid19
2020-03-20	NORMAJTORRES	Democrat	covid19
2020-03-17	REPJEFFDUNCAN	Republican	covid19

Patterns of variation over time

The table below details the first attestation of each referring expression in our **2020 Congressional Twitter corpus**. `coronavirus` hit the scene on 1-17, followed by `pandemic` on 1-22, `coronavirus pandemic` on 2-11, and `covid19` on 2-12 – the name for the disease coined by the World Health Organization on 2-11.

```
covid_tweets %>%
  group_by(covid_gram) %>%
  filter(date == min(date)) %>%
  arrange(date) %>%
  select(covid_gram, date, twitter) %>%
  knitr::kable()
```

covid_gram	date	twitter
coronavirus	2020-01-17	SENFEINSTEIN
pandemic	2020-01-22	MICHAELCBURGESS
pandemic	2020-01-22	SENTOMCOTTON
coronavirus pandemic	2020-02-11	SENATORHASSAN
covid19	2020-02-12	REPELIOTENGEL

Rates of reference to 2019 NOVEL CORONAVIRUS

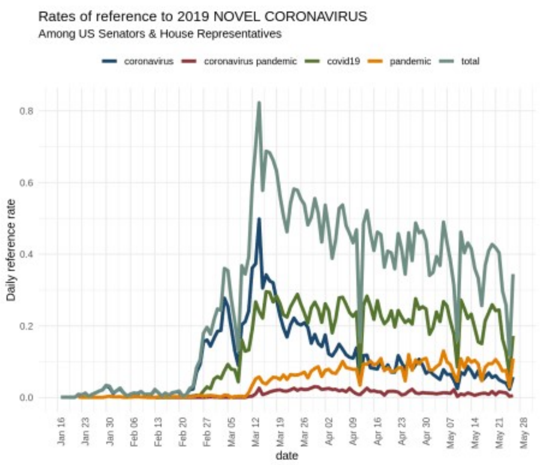
So, what lexical forms are Senators and House Reps using to reference `2019 NOVEL CORONAVIRUS` on Twitter? How often are they referring to `2019 NOVEL CORONAVIRUS`? And how have these patterns changed over time? To get a beat, we consider the daily rate of reference to `2019 NOVEL CORONAVIRUS`, and the rate at which each lexical variant has been used to reference `2019 NOVEL CORONAVIRUS`.

The **reference rate** for referring expression *X*, then, is approximated as the proportion of total tweets generated by members of Congress that contain referring expression *X*. The plot below illustrates daily rates of reference for each form from Jan 17 to March 27. Included is the total reference rate for `2019 NOVEL CORONAVIRUS`.

```
all <- covid_tweets %>%
  group_by(date) %>%
  summarize(n = n()) %>%
  left_join(all_tweets %>% group_by(date) %>% summarise(ts = sum(ts))) %>%
  mutate(per = n/ts,
         covid_gram = 'total') %>%
  select(date, covid_gram, n:per)


covid_tweets %>%
  group_by(date, covid_gram) %>% #,party,
  summarize(n = n()) %>%
  left_join(all_tweets %>% group_by(date) %>% summarise(ts = sum(ts))) %>%
  mutate(per = n/ts) %>%
  bind_rows(all) %>%

  ggplot() +
  geom_line(aes(x = date,
              y= per,
              color = covid_gram
              ), size = 1.5
          ) +
  theme_minimal() +
  ggthemes::scale_color_stata() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  scale_x_date(date_breaks = '2 days', date_labels = "%b %d") +
  theme(legend.position = 'top',
        legend.title = element_blank())  +
  ylab('Daily reference rate') +
  labs(title = 'Rates of reference to 2019 NOVEL CORONAVIRUS',
       subtitle = 'Among US Senators & House Representatives')
```

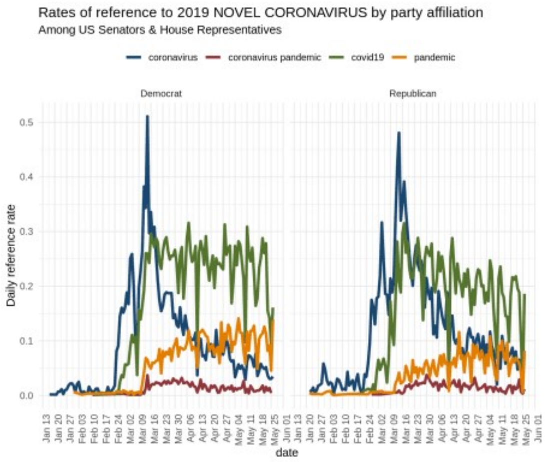


So, lots going on. Some `2019 NOVEL CORONAVIRUS` chatter through January and most of February. At Feb 24, we see a substantial jump. Reference to `2019 NOVEL CORONAVIRUS` as `coronvirus` has been most frequent since the onset; however, `covid19` has more recently taken the lead among members of Congress.

The plot below presents the same data disaggregated by party affiliation. Roughly the same profiles, which is interesting. With the exception of `covid19`, which shows an uptick with Dems not mirrored by Republicans. Likely just an anomaly.

```
covid_tweets %>%
  group_by(date, party, covid_gram) %>% #,party,
  summarize(n = n()) %>%
  left_join(all_tweets) %>%
  mutate(per = n/ts) %>%

  #filter(date < '2020-3-27') %>% # & date < '2020-3-27'
  filter(party != 'Independent') %>%
  ggplot() +
  geom_line(aes(x = date,
                y= per,
                color = covid_gram
                ),
            size = 1.25) +
  theme_minimal() +
  ggthemes::scale_color_stata() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        legend.title = element_blank(),
        legend.position = 'top')+
  scale_x_date(date_breaks = '1 week', date_labels = "%b %d") +
  facet_wrap(~party) + ylab('Daily reference rate') +
  labs(title = 'Rates of reference to 2019 NOVEL CORONAVIRUS by party affiliation',
       subtitle = 'Among US Senators & House Representatives')
```



Probability distributions

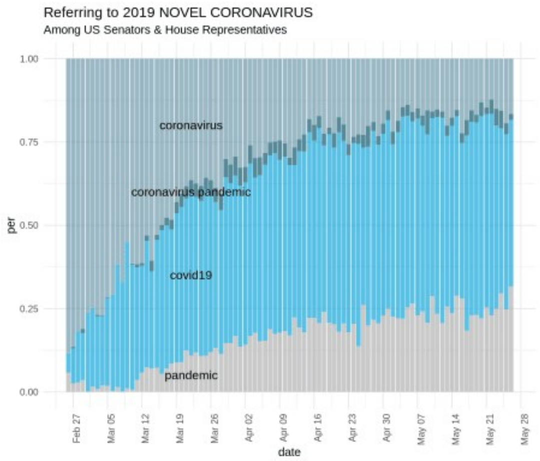
Lastly, we consider a proportional perspective on reference to `2019 NOVEL CORONAVIRUS`. Instead of total tweets, the denominator here becomes overall references to `2019 NOVEL CORONAVIRUS` on Twitter among members of Congress.

The figure below, then, illustrates daily probability distributions for forms used to reference `2019 NOVEL CORONAVIRUS`. `covid19` has slowly become the majority form on Twitter – `coronavirus` has become less and less prevalent. One explanation is that the effects of the virus in the US, ie, the disease, have become more prevalent and, hence, the proper use of the referring expression `covid19`. Another explanation is that `covid19` is shorter orthographically, and in the character-counting world of Twitter, a more efficient way to express the notion `2019 NOVEL CORONAVIRUS`. An empirical question for sure.

```
x1 <- covid_tweets %>%
  filter(date > '2020-2-25') %>%
  group_by(date, covid_gram) %>% #,party,
  summarize(n = n()) %>%
  mutate(per = n/sum(n))
```

```
x2 <- x1 %>%
  ggplot(aes(x=date, y=per, fill = covid_gram))+
  geom_bar(alpha = 0.65, stat = 'identity', width = .9) + #
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  theme(legend.position = "none")+
  ggthemes::scale_fill_economist() +
  scale_x_date(date_breaks = '1 day', date_labels = "%b %d") +
  labs(title = 'Referring to 2019 NOVEL CORONAVIRUS',
       subtitle = 'Among US Senators & House Representatives')

x2 +
    annotate(geom="text",
           x = c(rep(as.Date('2020-3-22'), 4)),
           y = c(.05, .35, .6, .8),
           label = c('pandemic', 'covid19', 'coronavirus pandemic', 'coronavirus'),
           size = 4, color = 'black')
```



Summary

So, a weekend & social distancing. Caveats galore, but for folks interested in language change & innovation & the establishment of convention in a community of speakers, something to keep an eye on.