CogSci Word Counts

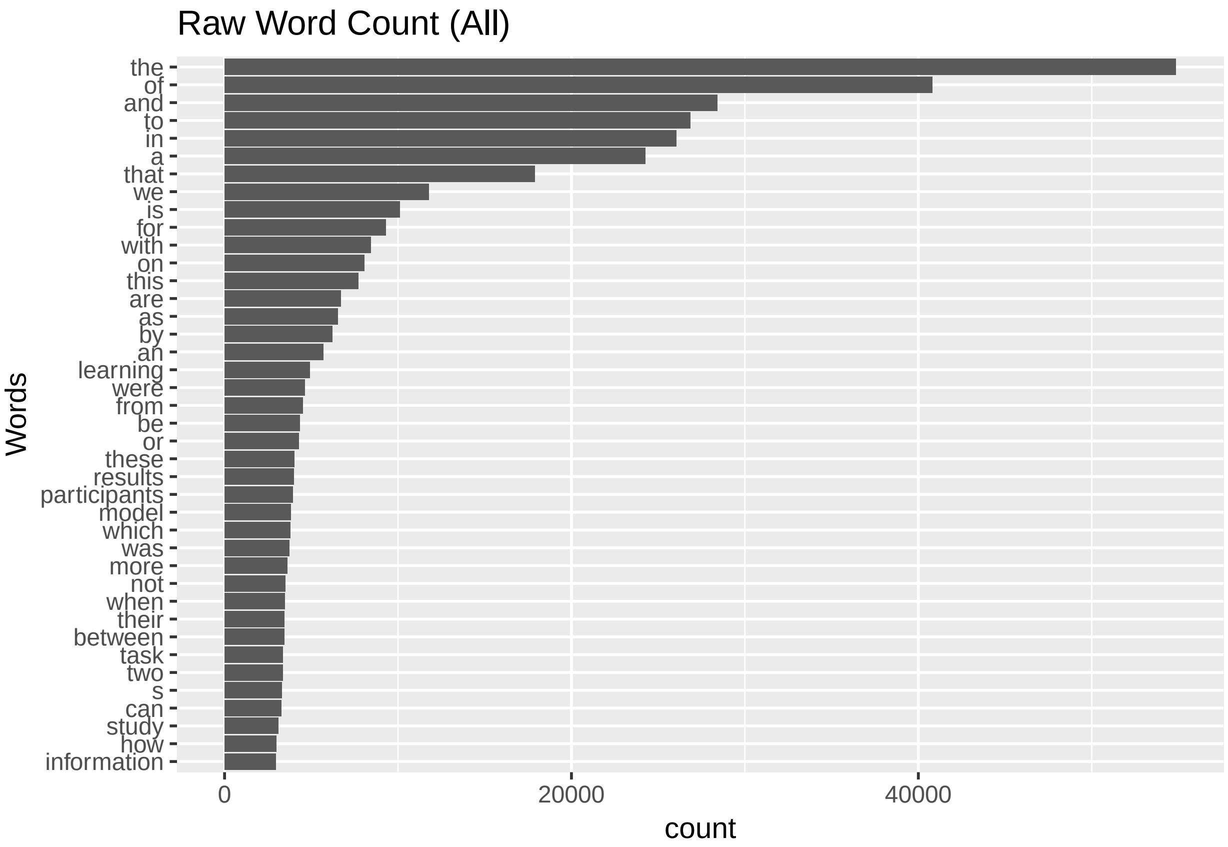
Lauren Oey

11/24/2018

```r
words <- df %>%
  mutate(abstract_cleaned = str_replace_all(abstract, c("e\\.g\\."="e1g1", "i\\.e\\."="i1e1")),
         abstract_cleaned = str_replace_all(abstract_cleaned, c("[^a-zA-Z0-9\\&\\s]"=" ", "[\\n\\t\\f]"=
         abstract_cleaned = str_replace_all(abstract_cleaned, c("e1g1"="e.g.", "i1e1"="i.e.")),
         word = strsplit(abstract_cleaned, " ")) %>%
  unnest(word) %>%
  filter(word != "") %>%
  mutate(lowerword = tolower(word))
```

Unfiltered Word Frequency Across All Abstracts

```r
words %>%
  group_by(lowerword) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  head(40) %>%
  ggplot(aes(x=reorder(lowerword,count), y=count)) +
  geom_bar(stat="identity") +
  ggtitle("Raw Word Count (All)") +
  scale_x_discrete("Words") +
  coord_flip()
```



```
ggsave("graphs/rawWordCount.png")
```

```
## Saving 6.5 x 4.5 in image
```

Gets Lemma & POS Info using CleanNLP; Removes Function Words (e.g. determiners, prepositions)

```
## First time running, you need to run this to extract the lemma/POS info from CleanNLP
## This takes awhile, so it's better to save the CSV file and read in the file for future use

#cnlp_init_udpipe()
#obj <- cnlp_annotate(df$abstract, as_strings = TRUE, doc_ids=df$title)
#obj_token <- cnlp_get_token(obj)
#write.csv(obj_token, "token_info.csv")
obj_token <- read.csv("token_info.csv")

## Filters out most POS, keeping nouns, verbs, adjectives, proper nouns, adverbs, numbers,
## and INTJ, which is kind of a mix of multiple things

obj_token_cleaned <- obj_token %>%
  filter(upos %in% c("NOUN", "VERB", "ADJ", "PROPN", "ADV", "NUM", "INTJ")) %>%
  mutate(title = id) %>%
  select("title","word","lemma")

write.csv(obj_token_cleaned, "token_info_cleaned.csv")
```

Joined CogSci Data + CleanNLP File (added lemma information)

```
words_full <- inner_join(words, unique(obj_token_cleaned), by=c("title","word")) %>%
  mutate(lowerlemma = tolower(lemma))
```

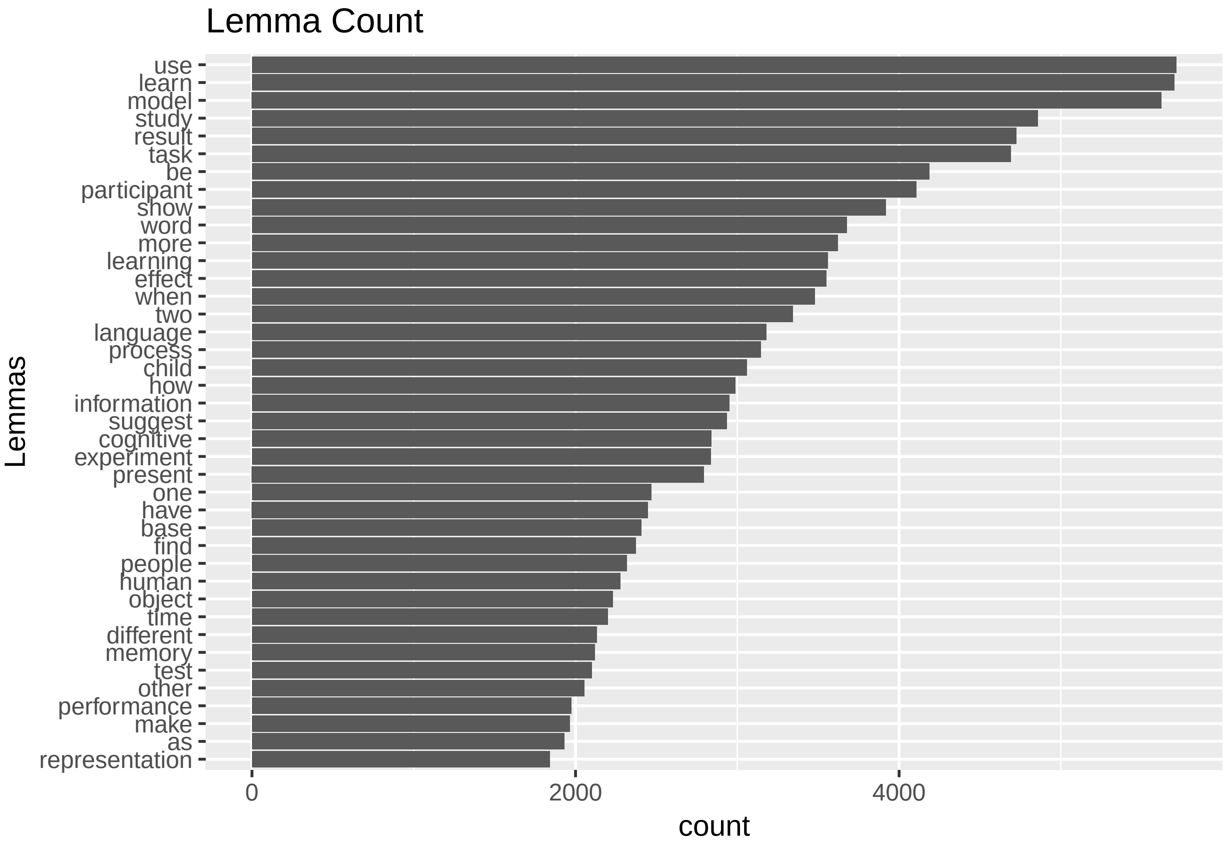
```
## Warning: Column `title` joining character vector and factor, coercing into
## character vector
```

```
## Warning: Column `word` joining character vector and factor, coercing into
## character vector
```

Lemma Word Count

```
wc_overall <- words_full %>%
  group_by(lowerlemma) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
write.csv(wc_overall, "wc_overall.csv")
```

```
ggplot(head(wc_overall, 40), aes(x=reorder(lowerlemma,count), y=count)) +
  geom_bar(stat="identity") +
  ggtitle("Lemma Count") +
  scale_x_discrete("Lemmas") +
  coord_flip()
```

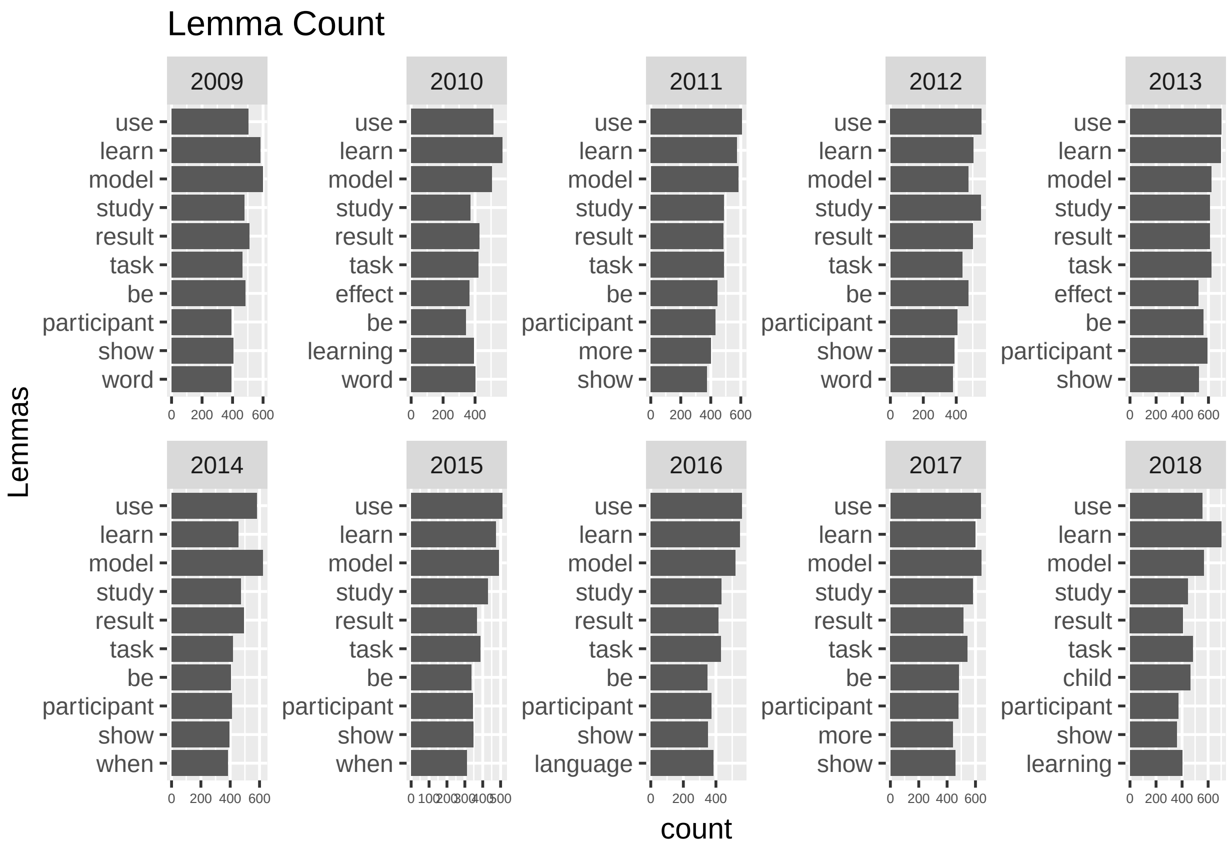


```
ggsave("graphs/overallPopWords.png")
```

```
## Saving 6.5 x 4.5 in image
```

```
wc_byYear <- words_full %>%
  group_by(year,lowerlemma) %>%
  summarise(count = n()) %>%
  arrange(year,desc(count))
write.csv(wc_byYear, "wc_byYear.csv")
```

```
wc_byYear %>%
  group_by(year) %>%
  top_n(10,count) %>%
  ggplot(aes(x=reorder(lowerlemma,count), y=count)) +
  geom_bar(stat="identity") +
  ggtitle("Lemma Count") +
  scale_x_discrete("Lemmas") +
  coord_flip() +
  facet_wrap(~year, nrow=2, scales="free") +
  theme(axis.text.x=element_text(size=5))
```



```
ggsave("graphs/wordsByYear.png")
```

```
## Saving 6.5 x 4.5 in image
```

```
wc_byTitle <- words_full %>%
  group_by(year,title,lowerlemma) %>%
  summarise(count = n()) %>%
  arrange(year,title,desc(count))
kable(head(wc_byTitle, 20))
```

year	title	lowerlemma	count
2009	‘If only’ counterfactuals and the exceptionality effect	exceptional	6
2009	‘If only’ counterfactuals and the exceptionality effect	outcome	6
2009	‘If only’ counterfactuals and the exceptionality effect	alternative	5
2009	‘If only’ counterfactuals and the exceptionality effect	lead	5
2009	‘If only’ counterfactuals and the exceptionality effect	when	4

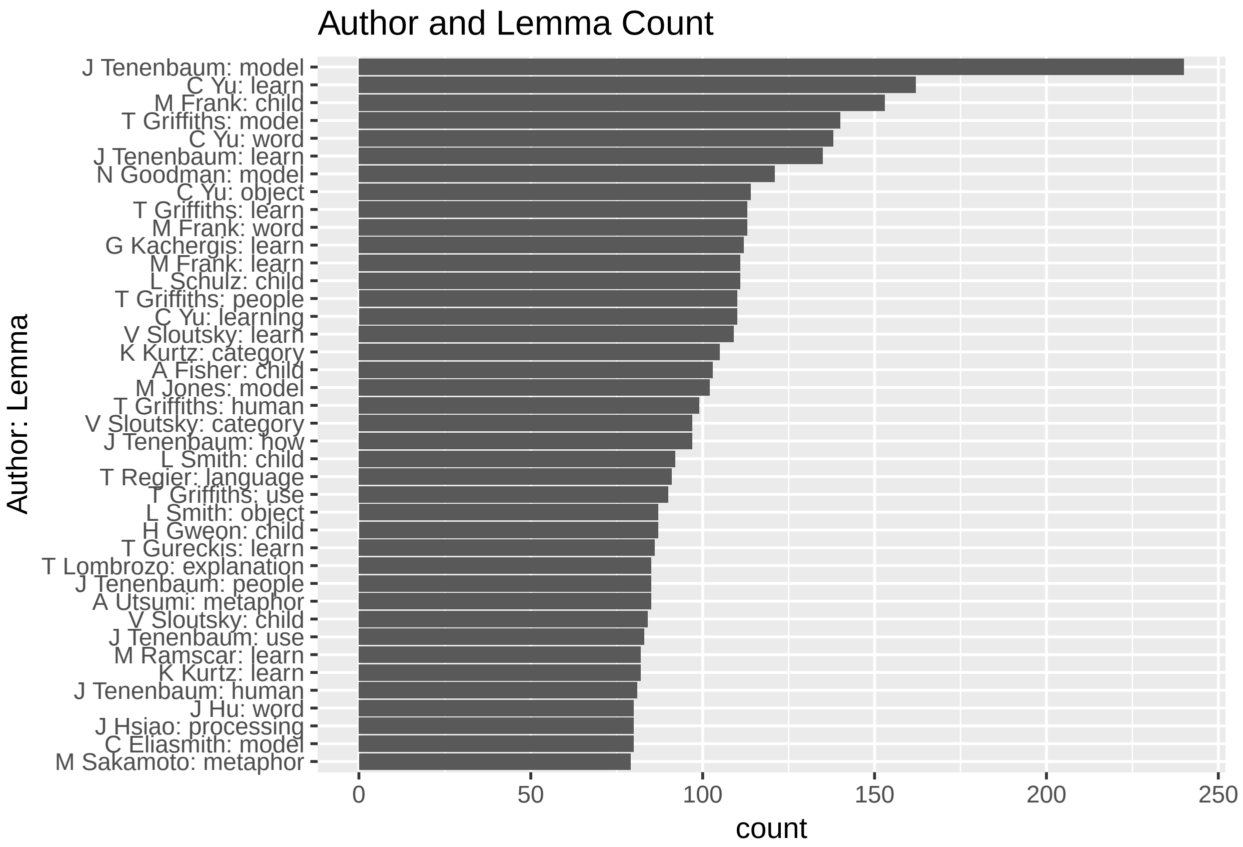
year	title	lowerlemma	count
2009	‘If only’ counterfactuals and the exceptionality effect	action	3
2009	‘If only’ counterfactuals and the exceptionality effect	better	3
2009	‘If only’ counterfactuals and the exceptionality effect	change	3
2009	‘If only’ counterfactuals and the exceptionality effect	counterfactual	3
2009	‘If only’ counterfactuals and the exceptionality effect	effect	3
2009	‘If only’ counterfactuals and the exceptionality effect	only	3
2009	‘If only’ counterfactuals and the exceptionality effect	also	2
2009	‘If only’ counterfactuals and the exceptionality effect	experiment	2
2009	‘If only’ counterfactuals and the exceptionality effect	experiments	2
2009	‘If only’ counterfactuals and the exceptionality effect	not	2
2009	‘If only’ counterfactuals and the exceptionality effect	reverse	2
2009	‘If only’ counterfactuals and the exceptionality effect	show	2
2009	‘If only’ counterfactuals and the exceptionality effect	usual	2
2009	‘If only’ counterfactuals and the exceptionality effect	1	1
2009	‘If only’ counterfactuals and the exceptionality effect	2	1

```r
write.csv(wc_byTitle, "wc_byTitle.csv")
```

```r
byAuthor <- words_full %>%
  mutate(authors=stri_trans_general(authors, "latin-ascii"),
         author=str_replace_all(authors, ",$", ""),
         author=str_replace_all(author, c(".*\n"="", " *\\(.*?\\)"="", "  "=" ")),
         author=str_replace_all(author, " & ", ", "),
         author=strsplit(author, ", ")) %>%
  unnest(author) %>%
  filter(!grepl("University|Institute|Center|Centre|Centro|School|Department|Dept|Unit|Hospital|\\d",
                author)) %>%
  mutate(author=trimws(author),
         authorWord = word(author, -1),
         authorAbbr = paste(substring(author,1,1), word(author, -1)))

wc_byAuthor <- byAuthor %>%
  group_by(authorAbbr, lowerlemma) %>%
  summarise(count = n()) %>%
  arrange(desc(count),authorAbbr)
write.csv(wc_byAuthor, "wc_byAuthor.csv")
```

```r
wc_byAuthor %>%
  mutate(authorLemma = paste(authorAbbr, lowerlemma, sep=": ")) %>%
  arrange(desc(count),authorAbbr) %>%
  head(40) %>%
  ggplot(aes(x=reorder(authorLemma,count), y=count)) +
  geom_bar(stat="identity") +
  ggtitle("Author and Lemma Count") +
  scale_x_discrete("Author: Lemma") +
  coord_flip()
```



```r
ggsave("graphs/AuthorWords.png")
```

```
## Saving 6.5 x 4.5 in image
```

```r
leaders <- byAuthor %>%
  select("year","authorAbbr","title") %>%
  unique() %>%
  group_by(year, authorAbbr) %>%
  summarise(totalPapers = n()) %>%
  top_n(3, totalPapers)
leaders %>% arrange(year,desc(totalPapers)) %>%
  kable(caption="Top 3 Authors w/ Most Papers by Year")
```

Table 2: Top 3 Authors w/ Most Papers by Year

year	authorAbbr	totalPapers
2009	T Griffiths	11
2009	J Tenenbaum	10
2009	M Lee	9
2010	L Smith	10
2010	J Tenenbaum	8
2010	T Griffiths	8
2011	T Griffiths	9
2011	W Fu	8
2011	E Krahmer	7
2011	G Storms	7
2011	J Tenenbaum	7
2011	R Dale	7

year	authorAbbr	totalPapers
2011	R Goldstone	7
2012	T Griffiths	12
2012	J Tenenbaum	11
2012	R Saxe	9
2013	R Dale	12
2013	C Eliasmith	8
2013	M Frank	8
2014	J Hu	11
2014	J Tenenbaum	11
2014	L Schulz	10
2014	N Goodman	10
2015	J Tenenbaum	15
2015	N Goodman	11
2015	M Richardson	10
2016	M Frank	12
2016	J Tenenbaum	10
2016	N Goodman	10
2017	J Tenenbaum	15
2017	T Griffiths	15
2017	K Smith	11
2018	J Tenenbaum	18
2018	T Griffiths	11
2018	M Frank	9

```r
## Approximately corresponds to top 3 authors by number of papers published
byAuthor %>%
  group_by(year, authorAbbr) %>%
  summarise(total = n()) %>%
  top_n(3, total) %>%
  arrange(year, desc(total)) %>%
  kable(caption="Top 3 Authors w/ Most Abstract Words by Year")
```

Table 3: Top 3 Authors w/ Most Abstract Words by Year

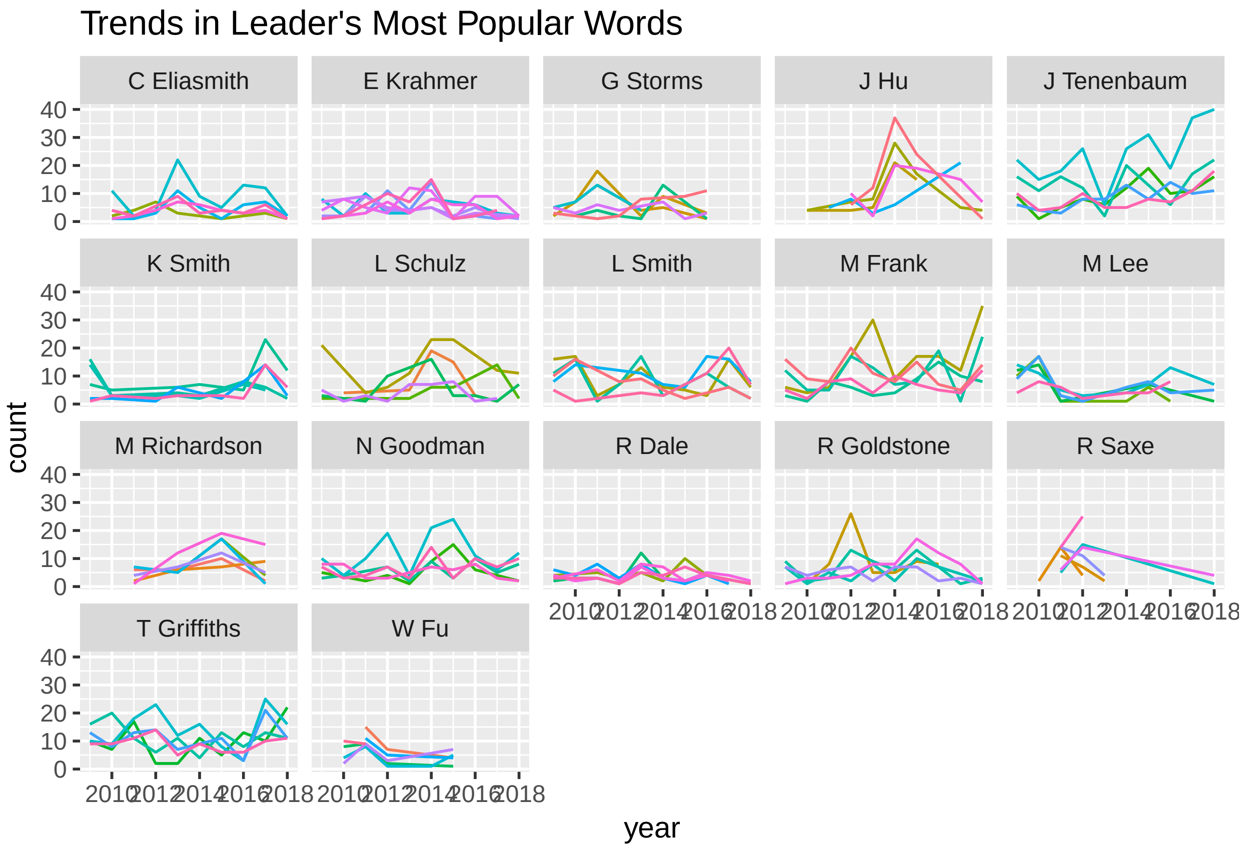
year	authorAbbr	total
2009	T Griffiths	920
2009	J Tenenbaum	849
2009	L Boroditsky	811
2010	L Smith	874
2010	M Lee	754
2010	R Shiffrin	687
2011	W Fu	736
2011	T Griffiths	730
2011	J Tenenbaum	642
2012	J Tenenbaum	1082
2012	T Griffiths	965
2012	R Saxe	891
2013	R Dale	1075
2013	I McLaren	914
2013	R Goldstone	712
2014	J Hu	1256

year	authorAbbr	total
2014	H Chen	1059
2014	J Tenenbaum	997
2015	J Tenenbaum	1277
2015	N Goodman	925
2015	M Richardson	914
2016	M Frank	968
2016	J Tenenbaum	874
2016	C Yu	777
2017	T Griffiths	1292
2017	J Tenenbaum	1280
2017	K Smith	921
2018	J Tenenbaum	1651
2018	T Griffiths	1067
2018	M Frank	990

```r
leaderFaveWords <- byAuthor %>%
  filter(authorAbbr %in% leaders$authorAbbr) %>%
  group_by(authorAbbr, lowerlemma) %>%
  summarise(count = n()) %>%
  top_n(5, count) %>%
  mutate(authorWord = paste(authorAbbr, lowerlemma, sep="_")) %>%
  arrange(authorAbbr, desc(count))
```

```r
wc_byAuthorYear <- byAuthor %>%
  group_by(year, authorAbbr, lowerlemma) %>%
  summarise(count = n()) %>%
  arrange(year, authorAbbr,desc(count))
write.csv(wc_byAuthorYear, "wc_byAuthorYear.csv")
```

```r
wc_byAuthorYear %>%
  filter(authorAbbr %in% unique(leaderFaveWords$authorAbbr)) %>%
  mutate(authorWord=paste(authorAbbr, lowerlemma, sep="_")) %>%
  filter(authorWord %in% unique(leaderFaveWords$authorWord)) %>%
  ggplot(aes(x=year, y=count, colour=lowerlemma)) +
  geom_line(stat="identity") +
  ggtitle("Trends in Leader's Most Popular Words") +
  facet_wrap(~authorAbbr) +
  guides(colour=FALSE)
```



```r
ggsave("graphs/leaderPopWords.png")
```

```
## Saving 6.5 x 4.5 in image
```