Fig3CD

```
source("source.R")
```

- path

```
path <- "data/dat_all.csv"
.bin <- 30
```

- import

```
dat_raw <-
  path %>% read_csv()
```

- count

```
dat_ngram <-
  dat_raw %>%
  arrange_ngram()
```

- arrange

```
dat_ngram_long <-
  dat_ngram %>%
  select(!data) %>%
  pivot_longer(cols = starts_with("gram"),
               names_to = "model",
               values_to = "data") %>%
  unnest(data) %>%
  rename(phrase = key) %>%
  group_by(type, Name, Stage, model, phrase) %>%
  summarise(n = mean(n)) %>%
  ungroup()

# output
write_csv(dat_ngram_long, "data/dat_ngram_long_ici.csv")
```

- JSD

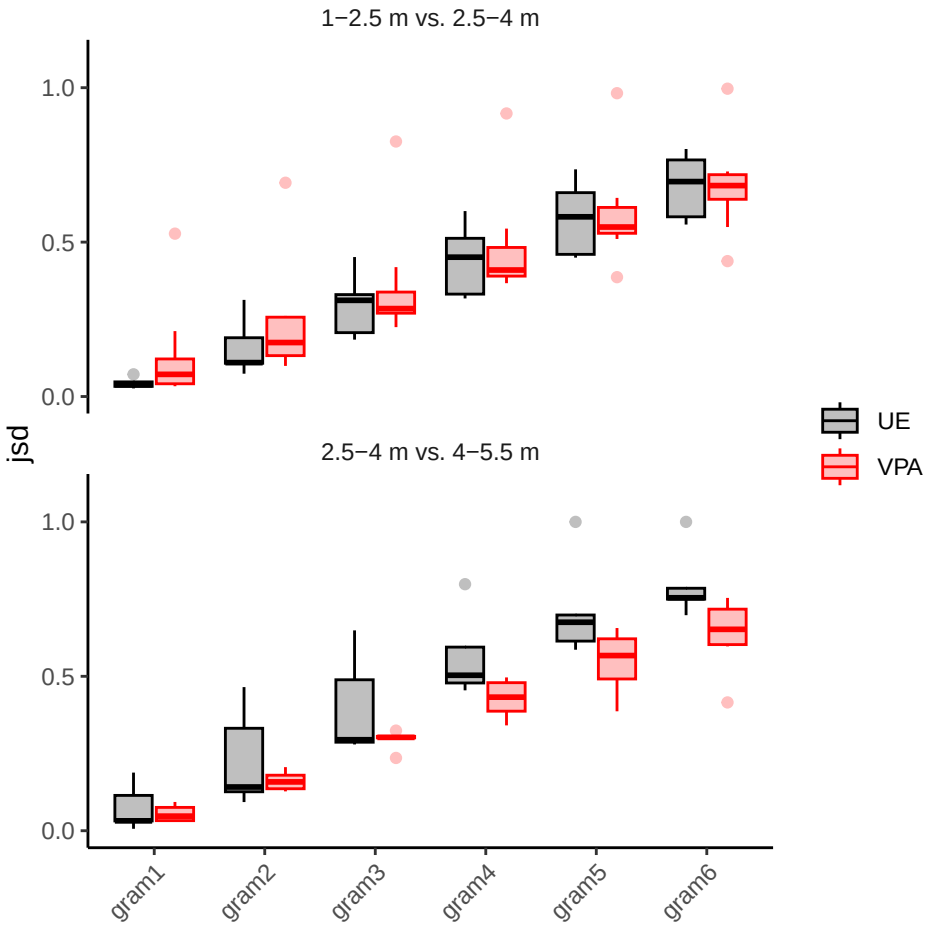
```
dat_ngram_rr_nest <-
  dat_ngram_long %>%
  group_nest(model) %>%
  mutate(rr = map(data, \(x){
    x %>%
```

```
      mutate(key = str_c(Stage, "_", Name)) %>%
      roundrobin::roundrobin(key, combination = TRUE) %>%
      separate(Var1, into = c("PM_1", "Name_1"), sep = "_")%>%
      separate(Var2, into = c("PM_2", "Name_2"), sep = "_") %>%
      filter(Name_1 == Name_2) %>%
      mutate(type = map_chr(data_Var1, ~.$type[[1]]))
  }))

dat_ngram_rr <-
  dat_ngram_rr_nest %>%
  select(!data) %>%
  unnest(rr) %>%
  mutate(data_Var1 = map(data_Var1, \(x){
    mutate(x, r1 = n / sum(n)) %>%
      select(phrase, r1)
  })) %>%
  mutate(data_Var2 = map(data_Var2, \(x){
    mutate(x, r2 = n / sum(n)) %>%
      select(phrase, r2)
  })) %>%
  mutate(for_jsd = map2(data_Var1, data_Var2, \(x, y){
    full_join(x, y, by = "phrase")
  } %>%
    mutate(r1 = if_else(is.na(r1), 0, r1)) %>%
    mutate(r2 = if_else(is.na(r2), 0, r2)) %>%
    select(r1, r2) %>%
    t()
  )) %>%
  mutate(jsd = map_dbl(for_jsd, philentropy::JSD))%>%
  mutate(kl = map_dbl(for_jsd, philentropy::KL))

dat_ngram_rr %>%
  select(!c(data_Var1, data_Var2))
```

```
## # A tibble: 204 x 9
##    model PM_1    Name_1     PM_2    Name_2     type  for_jsd           jsd    kl
##    <chr> <chr>   <chr>      <chr>   <chr>      <chr> <list>          <dbl> <dbl>
##  1 gram1 1-2.5 m Bach       2.5-4 m Bach       VPA   <dbl [2 x 10]> 0.0352 0.213
##  2 gram1 1-2.5 m Belarus    2.5-4 m Belarus    UE    <dbl [2 x 8]>  0.0474 0.235
##  3 gram1 1-2.5 m Belladonna 2.5-4 m Belladonna UE    <dbl [2 x 8]>  0.0381 0.188
##  4 gram1 1-2.5 m Brahms     2.5-4 m Brahms     VPA   <dbl [2 x 10]> 0.0530 0.482
##  5 gram1 1-2.5 m Camaro     2.5-4 m Camaro     VPA   <dbl [2 x 10]> 0.0433 0.276
##  6 gram1 1-2.5 m Cheecama   2.5-4 m Cheecama   VPA   <dbl [2 x 9]>  0.0916 0.348
##  7 gram1 1-2.5 m Chevrolet  2.5-4 m Chevrolet  VPA   <dbl [2 x 8]>  0.0334 0.126
##  8 gram1 1-2.5 m Chuppa     2.5-4 m Chuppa     UE    <dbl [2 x 9]>  0.0264 0.122
##  9 gram1 1-2.5 m Genge      2.5-4 m Genge      UE    <dbl [2 x 8]>  0.0719 0.378
## 10 gram1 1-2.5 m Ipsum      2.5-4 m Ipsum      VPA   <dbl [2 x 9]>  0.527  2.44
## # i 194 more rows
```

- visualization

```r
g_jsd <-
  dat_ngram_rr %>%
  mutate(key = str_c(PM_1, " vs. ", PM_2)) %>%
  filter(key != "1-2.5 m vs. 4-5.5 m") %>%
  ggplot() +
  aes(model, jsd) +
  geom_boxplot(aes(color = type, fill = type), alpha = 0.25) +
  scale_fill_manual(values = c(UE = "black", VPA = "red")) +
  scale_color_manual(values = c(UE = "black", VPA = "red")) +
  facet_wrap(~key, ncol = 1) +
  theme(axis.title.x = element_blank(),
        axis.text.x = element_text(angle = 45,
                                   hjust = 1)) +
  scale_y_continuous(breaks = c(0, 0.5, 1),
                     limits = c(0, 1.1))

g_jsd
```




```r
ggsave("fig/fig_3DC.png", width = 5, height = 5)

ggsave("fig/fig_3DC.svg", width = 5, height = 5)
```

- stat with Brunner-Munzel test

```
dat_stat <-
  dat_ngram_rr %>%
  mutate(key = str_c(PM_1, " vs. ", PM_2)) %>%
  filter(key != "1-2.5 m vs. 4-5.5 m") %>%
  group_nest(model, key, type) %>%
  pivot_wider(values_from = data,
              names_from = type) %>%
  mutate(stat = map2(UE, VPA, \(x, y){
    lawstat::brunner.munzel.test(x$jsd, y$jsd)
  })) %>%
  mutate(p = map_dbl(stat, ~.$p.value)) %>%
  arrange(key, model)

dat_stat %>% filter(p < 0.05) %>%  .$stat
```

```
## [[1]]
##
##  Brunner-Munzel Test
##
## data:  x$jsd and y$jsd
## Brunner-Munzel Test Statistic = -3.1238, df = 8.9437, p-value = 0.01234
## 95 percent confidence interval:
##  -0.1324509  0.3991176
## sample estimates:
## P(X<Y)+.5*P(X=Y)
##        0.1333333
##
##
## [[2]]
##
##  Brunner-Munzel Test
##
## data:  x$jsd and y$jsd
## Brunner-Munzel Test Statistic = -3.1238, df = 8.9437, p-value = 0.01234
## 95 percent confidence interval:
##  -0.1324509  0.3991176
## sample estimates:
## P(X<Y)+.5*P(X=Y)
##        0.1333333
##
##
## [[3]]
##
##  Brunner-Munzel Test
##
## data:  x$jsd and y$jsd
## Brunner-Munzel Test Statistic = -4.1906, df = 8.932, p-value = 0.002379
## 95 percent confidence interval:
##  -0.1161786  0.3161786
## sample estimates:
## P(X<Y)+.5*P(X=Y)
##              0.1
```

```r
sessionInfo()
```

```
## R version 4.3.3 (2024-02-29)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Sonoma 14.3
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapack.dylib;  LAPACK v
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: Asia/Tokyo
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lme4_1.1-35.5     Matrix_1.6-5      patchwork_1.2.0   data.table_1.15.0
##  [5] lubridate_1.9.3   forcats_1.0.0     stringr_1.5.1     dplyr_1.1.4
##  [9] purrr_1.0.2       readr_2.1.4       tidyr_1.3.0       tibble_3.2.1
## [13] ggplot2_3.5.1     tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] gtable_0.3.4      xfun_0.41         lattice_0.22-5    tzdb_0.4.0
##  [5] vctrs_0.6.5       tools_4.3.3       Rdpack_2.6        generics_0.1.3
##  [9] parallel_4.3.3    fansi_1.0.6       highr_0.10        pkgconfig_2.0.3
## [13] lifecycle_1.0.4   compiler_4.3.3    farver_2.1.1      textshaping_0.3.7
## [17] munsell_0.5.0     Kendall_2.2.1     htmltools_0.5.7   lawstat_3.6
## [21] yaml_2.3.7        pillar_1.9.0      nloptr_2.0.3      crayon_1.5.2
## [25] MASS_7.3-60.0.1   boot_1.3-29       nlme_3.1-164      tidyselect_1.2.0
## [29] digest_0.6.33     mvtnorm_1.2-4     stringi_1.8.3     splines_4.3.3
## [33] fastmap_1.1.1     grid_4.3.3        colorspace_2.1-0  cli_3.6.2
## [37] magrittr_2.0.3    utf8_1.2.4        withr_2.5.2       scales_1.3.0
## [41] bit64_4.0.5       timechange_0.2.0  rmarkdown_2.25    roundrobin_0.0.4
## [45] bit_4.0.5         ragg_1.2.6        hms_1.1.3         evaluate_0.23
## [49] knitr_1.45        rbibutils_2.2.16  philentropy_0.8.0 rlang_1.1.3
## [53] Rcpp_1.0.11       glue_1.7.0        svglite_2.1.2     rstudioapi_0.15.0
## [57] vroom_1.6.4       minqa_1.2.6       R6_2.5.1          systemfonts_1.0.5
```