

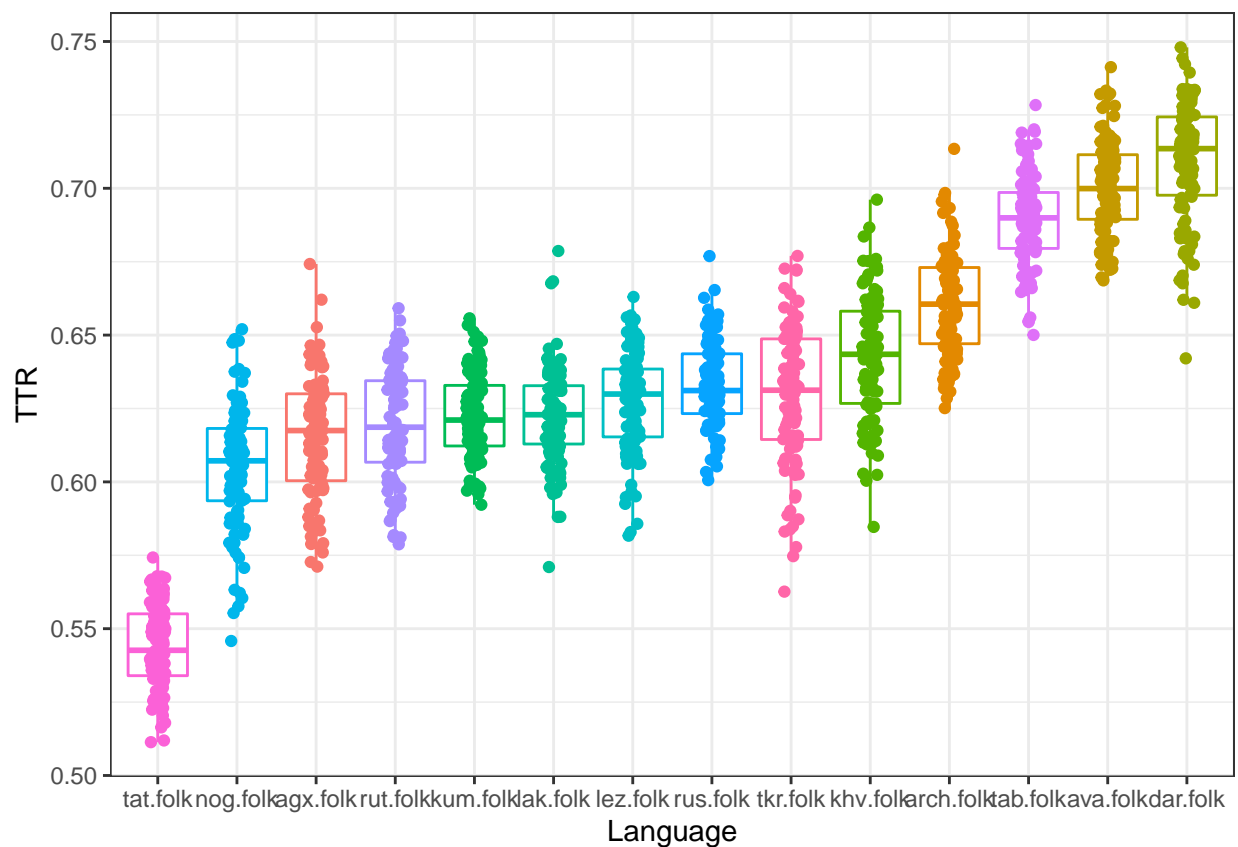
# TTR across languages of Daghestan

## 100 TTRs per language

```
df1 <- read.csv("TTR_100_datapoints_per_language.txt",  
               header = TRUE, sep = ",")  
df1_long <- gather(df1)  
head(df1_long)
```

```
##      key      value  
## 1 agx.folk 0.5849057  
## 2 agx.folk 0.6420846  
## 3 agx.folk 0.6250000  
## 4 agx.folk 0.6620553  
## 5 agx.folk 0.6105675  
## 6 agx.folk 0.5928144
```

```
df1_long %>%  
  ggplot(aes(reorder(key, value, FUN = median), value, color = key))+  
  geom_boxplot(outlier.shape = NA)+  
  geom_jitter(width = 0.1)+  
  labs(x = "Language",  
       y = "TTR")+  
  theme_bw()+  
  theme(legend.position = "none")
```



## Pairwise t-test

Holm-Bonferroni method

```
p.values <- pairwise.t.test(df1_long$value, df1_long$key,
                             paired = T, p.adjust.method = "holm")
p.values
```

```
##
## Pairwise comparisons using paired t tests
##
## data: df1_long$value and df1_long$key
##
##      agx.folk arch.folk ava.folk dar.folk khv.folk kum.folk lak.folk
## arch.folk < 2e-16 - - - - -
## ava.folk < 2e-16 < 2e-16 - - - -
## dar.folk < 2e-16 < 2e-16 0.0115 - - -
## khv.folk 6.2e-14 2.0e-07 < 2e-16 < 2e-16 - - -
## kum.folk 0.0929 < 2e-16 < 2e-16 < 2e-16 8.1e-10 - -
## lak.folk 0.1310 < 2e-16 < 2e-16 < 2e-16 1.6e-08 1.0000 -
## lez.folk 0.0011 < 2e-16 < 2e-16 < 2e-16 3.8e-06 0.1695 0.2973
## nog.folk 0.0097 < 2e-16 < 2e-16 < 2e-16 < 2e-16 5.9e-08 2.9e-08
## rus.folk 1.1e-09 < 2e-16 < 2e-16 < 2e-16 0.0060 2.2e-05 8.0e-05
## rut.folk 1.0000 < 2e-16 < 2e-16 < 2e-16 1.6e-11 1.0000 1.0000
```

```
## tab.folk < 2e-16 < 2e-16 4.1e-05 8.2e-09 < 2e-16 < 2e-16 < 2e-16
## tat.folk < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16
## tkr.folk 9.0e-05 7.1e-16 < 2e-16 < 2e-16 0.0013 0.1310 0.2364
##      lez.folk nog.folk rus.folk rut.folk tab.folk tat.folk
## arch.folk - - - - - -
## ava.folk - - - - - -
## dar.folk - - - - - -
## khv.folk - - - - - -
## kum.folk - - - - - -
## lak.folk - - - - - -
## lez.folk - - - - - -
## nog.folk 1.6e-13 - - - - -
## rus.folk 0.1695 < 2e-16 - - - -
## rut.folk 0.0283 3.9e-05 1.6e-06 - - -
## tab.folk < 2e-16 < 2e-16 < 2e-16 < 2e-16 - -
## tat.folk < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 -
## tkr.folk 1.0000 6.7e-12 1.0000 0.0157 < 2e-16 < 2e-16
##
## P value adjustment method: holm
```

## Effect sizes

```
library("rstatix")
```

```
##
## Attaching package: 'rstatix'
```

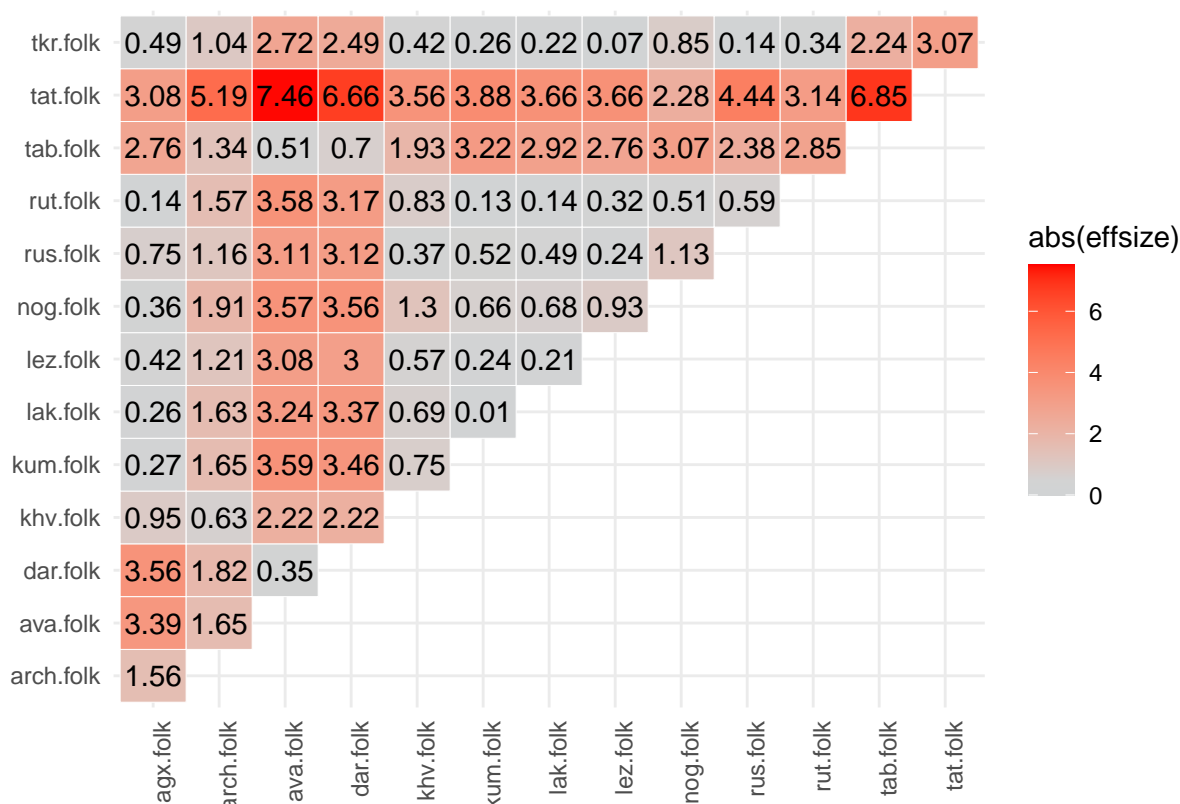
```
## The following object is masked from 'package:stats':
##
##      filter
```

```
effect.sizes <- cohens_d(df1_long, value ~ key, paired = T)
print(effect.sizes)
```

```
## # A tibble: 91 x 7
##   .y. group1 group2 effsize n1 n2 magnitude
## * <chr> <chr> <chr> <dbl> <int> <int> <ord>
## 1 value agx.folk arch.folk -1.56 100 100 large
## 2 value agx.folk ava.folk -3.39 100 100 large
## 3 value agx.folk dar.folk -3.56 100 100 large
## 4 value agx.folk khv.folk -0.946 100 100 large
## 5 value agx.folk kum.folk -0.275 100 100 small
## 6 value agx.folk lak.folk -0.257 100 100 small
## 7 value agx.folk lez.folk -0.421 100 100 small
## 8 value agx.folk nog.folk 0.356 100 100 small
## 9 value agx.folk rus.folk -0.747 100 100 moderate
## 10 value agx.folk rut.folk -0.136 100 100 negligible
## # ... with 81 more rows
```

## Effect size heatmap

```
effect.sizes.plot <- ggplot(as.data.frame(effect.sizes), aes(group1, group2)) +
  geom_tile(aes(fill = abs(effsize)), color = "white") +
  scale_fill_gradient2(low = "light blue", mid = "light grey", high = "red",
    midpoint = 0.5, limit = c(0, 7.5)) +
  geom_text(aes(label = round(abs(effsize), 2))) +
  labs(x = "", y = "") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
effect.sizes.plot
```



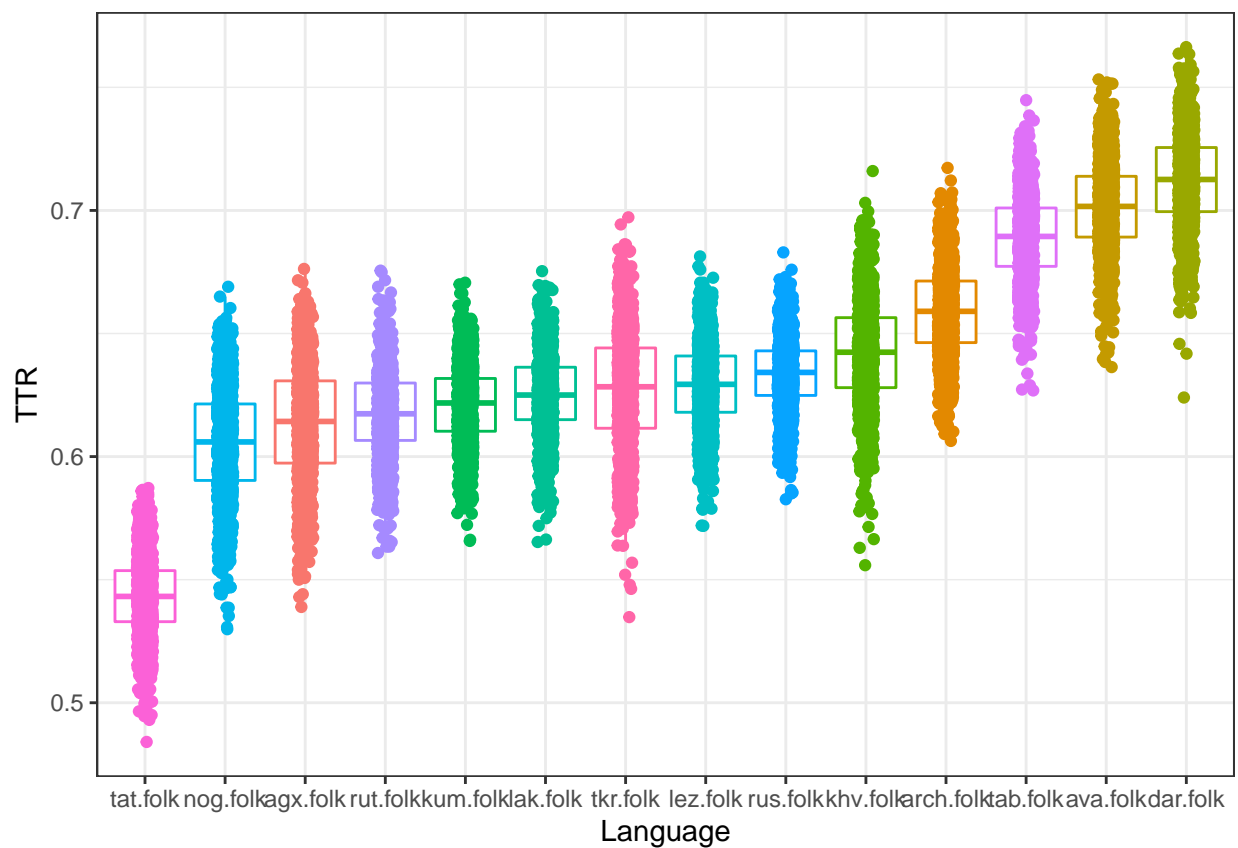
## 1000 TTRs per language

```
df2 <- read.csv("TTR_1000_datapoints_per_language.txt",
  header = TRUE, sep = ",")
df2_long <- gather(df2)
head(df2_long)
```

```
##      key      value
## 1 agx.folk 0.5940594
```

```
## 2 agx.folk 0.6343656
## 3 agx.folk 0.6179104
## 4 agx.folk 0.6000000
## 5 agx.folk 0.6305419
## 6 agx.folk 0.6510469
```

```
df2_long %>%
  ggplot(aes(reorder(key, value, FUN = median), value, color = key))+
  geom_boxplot(outlier.shape = NA)+
  geom_jitter(width = 0.1)+
  labs(x = "Language",
       y = "TTR")+
  theme_bw()+
  theme(legend.position = "none")
```



## Pairwise t-test

Holm-Bonferroni method

```
p.values <- pairwise.t.test(df2_long$value, df2_long$key,
                             paired = T, p.adjust.method = "holm")
p.values
```

```
##
```

```
## Pairwise comparisons using paired t tests
##
## data: df2_long$value and df2_long$key
##
##      agx.folk arch.folk ava.folk dar.folk khv.folk kum.folk lak.folk
## arch.folk < 2e-16 - - - - -
## ava.folk < 2e-16 < 2e-16 - - - - -
## dar.folk < 2e-16 < 2e-16 < 2e-16 - - - - -
## khv.folk < 2e-16 < 2e-16 < 2e-16 < 2e-16 - - - - -
## kum.folk 6.2e-15 < 2e-16 < 2e-16 < 2e-16 < 2e-16 - - - - -
## lak.folk < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 5.0e-07 -
## lez.folk < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 3.3e-07
## nog.folk 6.1e-15 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16
## rus.folk < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16
## rut.folk 3.3e-05 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 4.9e-05 < 2e-16
## tab.folk < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16
## tat.folk < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16
## tkr.folk < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 1.0e-11 0.0067
##      lez.folk nog.folk rus.folk rut.folk tab.folk tat.folk
## arch.folk - - - - -
## ava.folk - - - - -
## dar.folk - - - - -
## khv.folk - - - - -
## kum.folk - - - - -
## lak.folk - - - - -
## lez.folk - - - - -
## nog.folk < 2e-16 - - - - -
## rus.folk 3.6e-08 < 2e-16 - - - - -
## rut.folk < 2e-16 < 2e-16 < 2e-16 - - - - -
## tab.folk < 2e-16 < 2e-16 < 2e-16 < 2e-16 - - - - -
## tat.folk < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 - - - - -
## tkr.folk 0.1477 < 2e-16 2.1e-09 < 2e-16 < 2e-16 < 2e-16
##
## P value adjustment method: holm
```

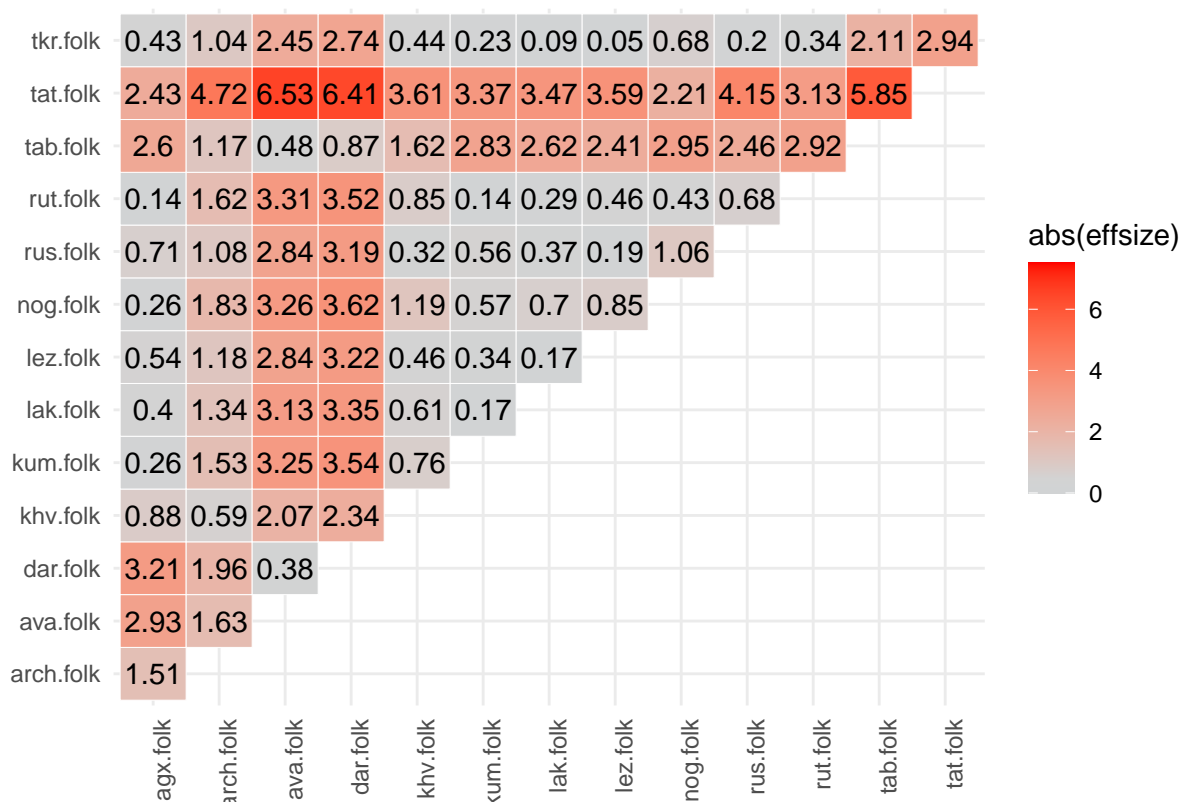
## Effect sizes

```
effect.sizes <- cohens_d(df2_long, value ~ key, paired = T)
head(effect.sizes)
```

```
## # A tibble: 6 x 7
##   .y. group1 group2 effsize n1 n2 magnitude
##   <chr> <chr> <chr> <dbl> <int> <int> <ord>
## 1 value agx.folk arch.folk -1.51 1000 1000 large
## 2 value agx.folk ava.folk -2.93 1000 1000 large
## 3 value agx.folk dar.folk -3.21 1000 1000 large
## 4 value agx.folk khv.folk -0.878 1000 1000 large
## 5 value agx.folk kum.folk -0.260 1000 1000 small
## 6 value agx.folk lak.folk -0.405 1000 1000 small
```

## Effect size heatmap

```
effect.sizes.plot <- ggplot(as.data.frame(effect.sizes), aes(group1, group2)) +
  geom_tile(aes(fill = abs(effsize)), color = "white") +
  scale_fill_gradient2(low = "light blue", mid = "light grey", high = "red",
    midpoint = 0.5, limit = c(0, 7.5)) +
  geom_text(aes(label = round(abs(effsize), 2))) +
  labs(x = "", y = "") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
effect.sizes.plot
```



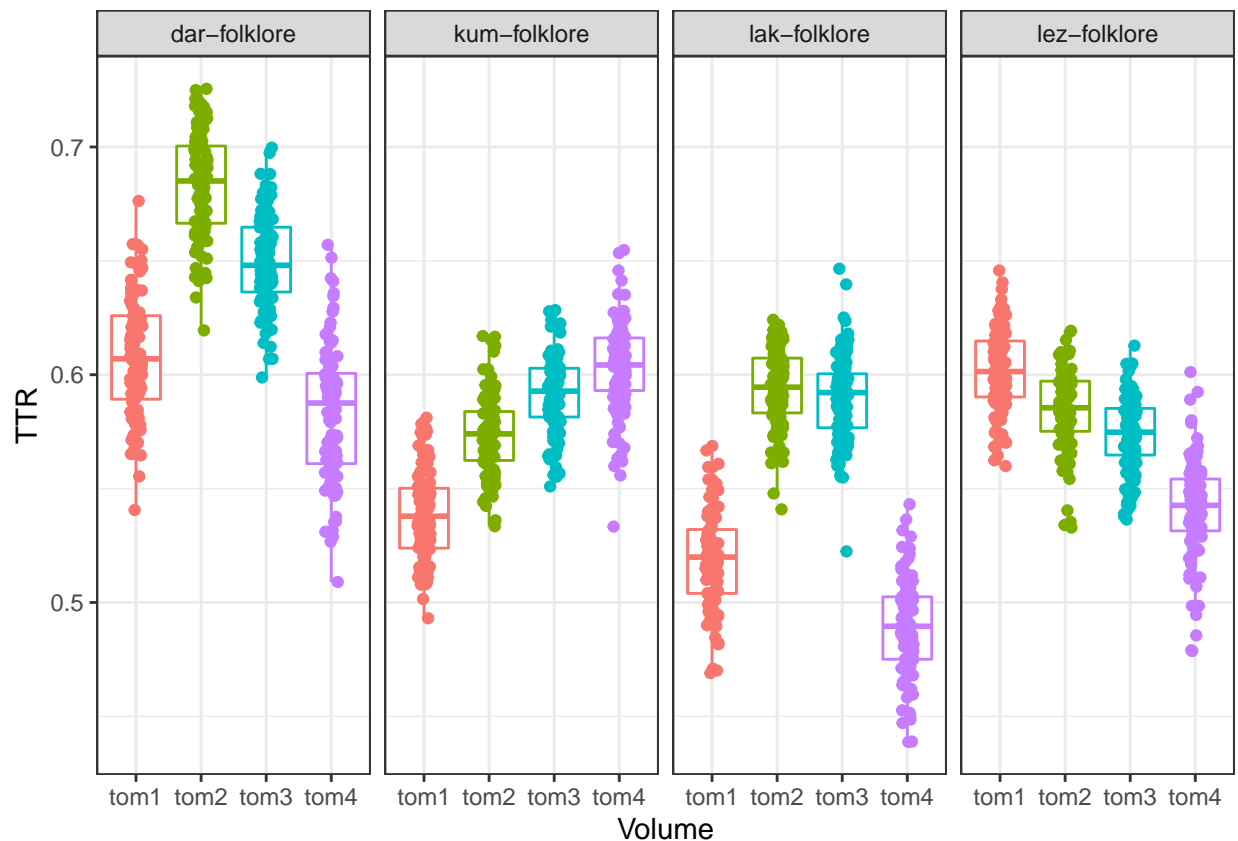
## Volume comparison

```
df3 <- read.csv("volume_comparison_100_datapoints_per_volume.txt",
  header = TRUE, sep = ",")
head(df3)
```

```
##   Volume      TTR    Language
## 1   tom1 0.5014985 kum-folklore
## 2   tom1 0.5276680 kum-folklore
```

```
## 3 tom1 0.5345850 kum-folklore
## 4 tom1 0.5811209 kum-folklore
## 5 tom1 0.5374016 kum-folklore
## 6 tom1 0.5351137 kum-folklore
```

```
df3 %>%
  ggplot(aes(Volume, TTR, color = Volume))+
  geom_boxplot(outlier.shape = NA)+
  geom_jitter(width = 0.1)+
  labs(x = "Volume",
       y = "TTR")+
  theme_bw()+
  theme(legend.position = "none")+
  facet_grid(cols = vars(Language))
```



## Pairwise t-tests

Holm-Bonferroni method

```
p.values <- pairwise.t.test(filter(df3, Language == "dar-folklore")$TTR,
                             filter(df3, Language == "dar-folklore")$Volume,
                             paired = T, p.adjust.method = "holm")
p.values
```



```
##
## Pairwise comparisons using paired t tests
##
## data: filter(df3, Language == "dar-folklore")$TTR and filter(df3, Language == "dar-folklore")$Volume
##
##      tom1      tom2      tom3
## tom2 < 2e-16 -      -
## tom3 < 2e-16 < 2e-16 -
## tom4 6.1e-08 < 2e-16 < 2e-16
##
## P value adjustment method: holm
```

```
p.values <- pairwise.t.test(filter(df3, Language == "kum-folklore")$TTR,
                             filter(df3, Language == "kum-folklore")$Volume,
                             paired = T, p.adjust.method = "holm")
p.values
```

```
##
## Pairwise comparisons using paired t tests
##
## data: filter(df3, Language == "kum-folklore")$TTR and filter(df3, Language == "kum-folklore")$Volume
##
##      tom1      tom2      tom3
## tom2 < 2e-16 -      -
## tom3 < 2e-16 1.6e-12 -
## tom4 < 2e-16 < 2e-16 2.5e-05
##
## P value adjustment method: holm
```

```
p.values <- pairwise.t.test(filter(df3, Language == "lak-folklore")$TTR,
                             filter(df3, Language == "lak-folklore")$Volume,
                             paired = T, p.adjust.method = "holm")
p.values
```

```
##
## Pairwise comparisons using paired t tests
##
## data: filter(df3, Language == "lak-folklore")$TTR and filter(df3, Language == "lak-folklore")$Volume
##
##      tom1      tom2      tom3
## tom2 <2e-16 -      -
## tom3 <2e-16 0.14   -
## tom4 <2e-16 <2e-16 <2e-16
##
## P value adjustment method: holm
```

```
p.values <- pairwise.t.test(filter(df3, Language == "lez-folklore")$TTR,
                             filter(df3, Language == "lak-folklore")$Volume,
                             paired = T, p.adjust.method = "holm")
p.values
```

```
##
```

```

## Pairwise comparisons using paired t tests
##
## data: filter(df3, Language == "lez-folklore")$TTR and filter(df3, Language == "lak-folklore")$Volum
##
##      tom1      tom2      tom3
## tom2 9.6e-09 -      -
## tom3 < 2e-16 2.5e-05 -
## tom4 < 2e-16 < 2e-16 < 2e-16
##
## P value adjustment method: holm

```