Entropy Analyses for Characters and White-Space-Separated Strings

Chris Bentz

22/02/2021

Load libraries

If the libraries are not installed yet, you need to install them using, for example, the command: install.packages("ggplot2"). For the Hrate package this is different, since it comes from github. The devtools library needs to be installed, and then the install_github() function is used.

```
library(stringr)
library(ggplot2)
library(ggrepel)
library(plyr)
library(ggExtra)
library(ggpubr)
```

```
##
## Attaching package: 'ggpubr'

## The following object is masked from 'package:plyr':
##
##     mutate
```

Load Data

Load data table with quantitative measures per text file.

```
# load estimations from stringBase corpus
estimations.df.sb <- read.csv("/home/chris/Github/StringBase/code/Tables/output_stringBase.csv")
head(estimations.df.sb)
```

```
##              filename subcorpus     code huni.chars hrate.chars huni.strings
## 1 animal_bhg_0001.txt    animal bhg_0001   3.494751    2.774313     4.017922
## 2 animal_bhg_0002.txt    animal bhg_0002   2.988396    2.825957     4.486239
## 3 animal_bhg_0003.txt    animal bhg_0003   3.471783    2.922633     4.330952
## 4 animal_bhg_0004.txt    animal bhg_0004   3.061043    2.682918     4.005791
## 5 animal_bhg_0005.txt    animal bhg_0005   3.464148    2.950704     4.297151
## 6 animal_bhg_0006.txt    animal bhg_0006   2.554254    2.730865     4.355434
##   hrate.strings ttr.chars ttr.strings   rm.chars
## 1      1.512035      0.18   0.1800000 0.00000000
## 2      2.035147      0.18   0.2637363 0.01219512
## 3      2.171960      0.21   0.2400000 0.00000000
## 4      2.378732      0.15   0.2300000 0.01176471
## 5      2.623054      0.21   0.4098361 0.01265823
## 6      2.428962      0.15   0.2600000 0.07058824
```

```
# load estimations from 100LC corpus
estimations.df.100lc <- read.csv("/home/chris/Github/StringBase/code/Tables/output_100LC_1Ksample.csv")
head(estimations.df.100lc)
```

```
##           filename subcorpus         code huni.chars hrate.chars huni.strings
## 1   tur_nfi_54.txt       tur   tur_nfi_54   4.275954    2.902026     6.503856
## 2  cmn_nfi_274.txt       cmn  cmn_nfi_274   6.117577    4.599791     4.772186
## 3 heb_nfi_1043.txt       heb heb_nfi_1043   4.546442    3.224313     6.376307
## 4  fra_fic_103.txt       fra  fra_fic_103   3.857725    2.116692     5.910015
## 5   ell_fic_49.txt       ell   ell_fic_49   4.766503    3.222067     6.165113
## 6  eus_nfi_630.txt       eus  eus_nfi_630   4.010809    2.708890     6.516307
##   hrate.strings ttr.chars ttr.strings   rm.chars
## 1      5.069336 0.3103448   0.9300000 0.00000000
## 2      2.875651 0.8241758   0.7073171 0.12500000
## 3      4.773364 0.3604651   0.8700000 0.03636364
## 4      4.328980 0.2209302   0.7000000 0.02985075
## 5      4.638224 0.4047619   0.8100000 0.00000000
## 6      4.993779 0.2500000   0.9400000 0.01449275
```

Add meta-information to 100LC (if needed).

```
# load meta-info from 100LC
meta.info <- read.csv("/home/chris/Data/100LC_Dumps/csv/file.csv")
meta.info <- meta.info[, 1:12]
# merge with estimations
estimations.df.100lc.meta <- merge(estimations.df.100lc, meta.info, by = "filename")
```

Combine 100LC and stringBase estimations.

```
# change labels in column ``subcorpus'' for 100LC (otherwise there are too many to plot)
estimations.df.100lc$subcorpus <- rep("writing", nrow(estimations.df.100lc))
estimations.df.combined <- rbind(estimations.df.100lc, estimations.df.sb)
```

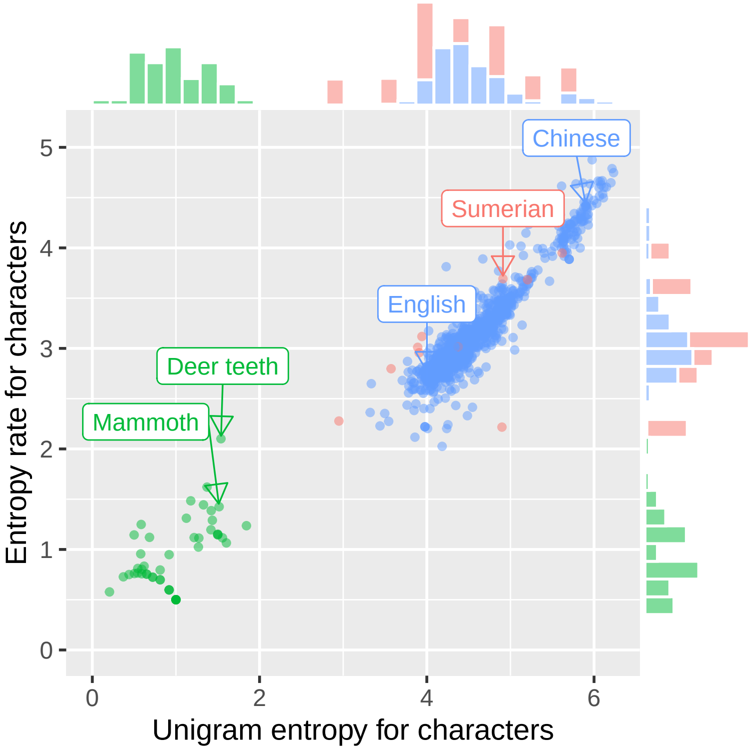
Select subcorpora (if needed).

```
selected <- c("writing", "ancient", "paleolithic")
estimations.df.combined <- estimations.df.combined[estimations.df.combined$subcorpus %in% selected, ]
```

Scatterplots

Entropy rate vs. unigram entropy for characters

```
# plot
huni.hrate.chars.plot <- ggplot(estimations.df.combined,
                                aes(x = huni.chars, y = hrate.chars,
                                    colour = subcorpus)) +
  geom_point(alpha = 0.5, size  = 1) +
  # geom_smooth(method = "lm") +
  xlim(0, max(estimations.df.combined$huni.chars)) +
  ylim(0, max(estimations.df.combined$hrate.chars)) +
  #theme(legend.position = "bottom") +
  #geom_rug() +
  #geom_segment(x = 0, y = 0, xend = 10, yend = 10, colour = "black",
            #linetype = "dashed", size = 0.3) +
  #geom_text(hjust = 0, nudge_x = 0, size = 2) +
  #geom_label_repel(aes(label = code), force = 0.5, force.pull = 5, label.size = 0.5, size = 3) +
  geom_label_repel(data = estimations.df.combined[estimations.df.combined$code == "sgr_0001" |
                                                   estimations.df.combined$code == "vhc_0145" |
                                                   estimations.df.combined$code == "sum_0003" |
                                                   estimations.df.combined$code == "cmn_0001" |
                                                   estimations.df.combined$code == "eng_0001" ,],
                    label = c("Deer teeth", "Mammoth", "Sumerian", "Chinese", "English"),
                    size = 3, arrow = arrow(length = unit(0.04, "npc"),
                              type = "closed", ends = "last"), nudge_y = 0.7,
                              segment.size  = 0.3) +
  labs(x = "Unigram entropy for characters", y = "Entropy rate for characters") +
  theme(legend.position = "none")
huni.hrate.chars.plot <- ggMarginal(huni.hrate.chars.plot, groupFill = T, type = "histogram", colour =
huni.hrate.chars.plot
```

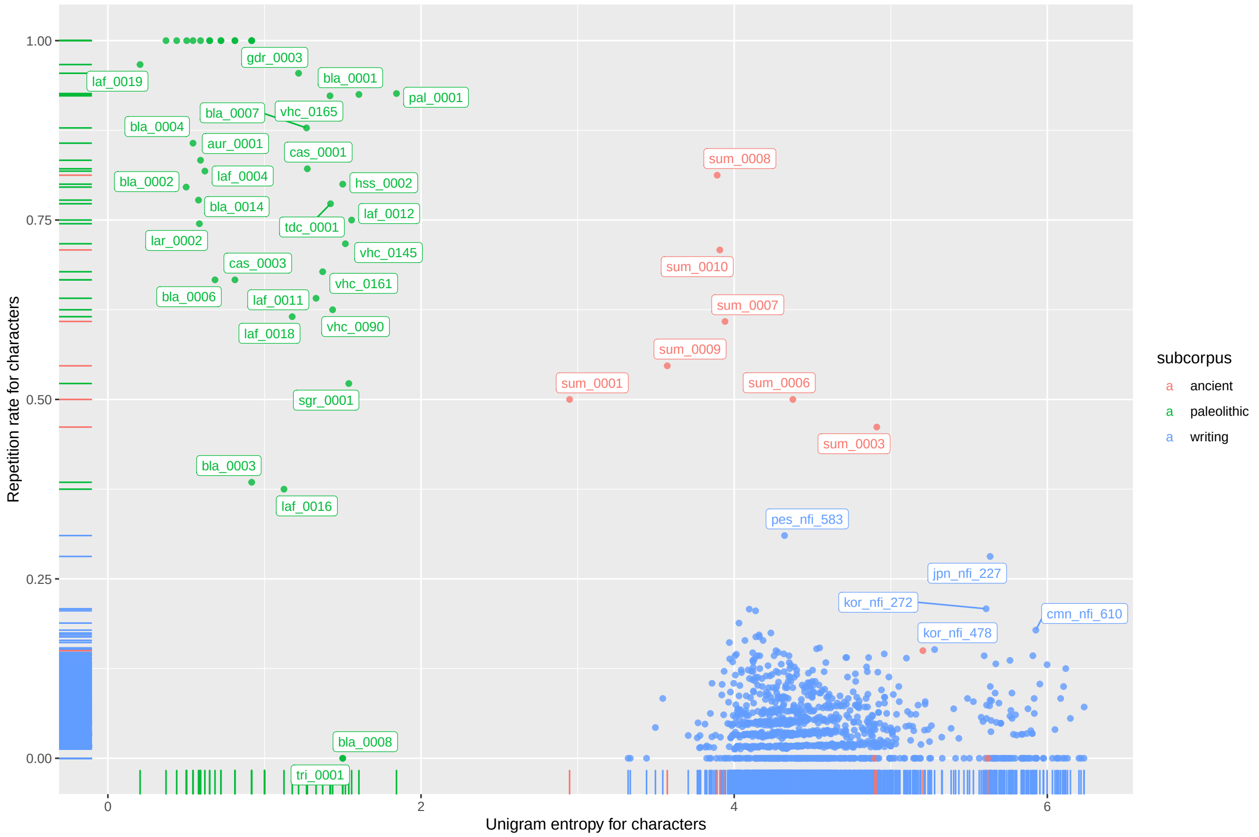


Safe complete figure to file

```
ggsave("Figures/huni_hrate_chars.pdf", huni.hrate.chars.plot, dpi = 300, width = 4, height = 4, device =
```

Unigram entropy vs repetition rate for characters

```
huni.rm.chars.plot <- ggplot(estimations.df.combined,
                              aes(x = huni.chars, y = rm.chars, colour = subcorpus)) +
  geom_point(alpha = 0.8, size  = 1.5) +
  #geom_smooth(method = "lm") +
  xlim(0, max(estimations.df.combined$huni.chars)) +
  ylim(0, max(estimations.df.combined$rm.chars)) +
  #theme(legend.position = "bottom") +
  geom_rug() +
  #geom_text(hjust = 0, nudge_x = 0.1, size = 2) +
  geom_label_repel(aes(label = code), force = 0.5, force.pull = 5, label.size = 0.2, size = 3) +
  labs(x = "Unigram entropy for characters", y = "Repetition rate for characters")
huni.rm.chars.plot
```



Safe complete figure to file

```
ggsave("Figures/huni_rm_chars.pdf", huni.rm.chars.plot, width = 13, height = 8, dpi = 300,
       scale = 1, device = cairo_pdf)
```

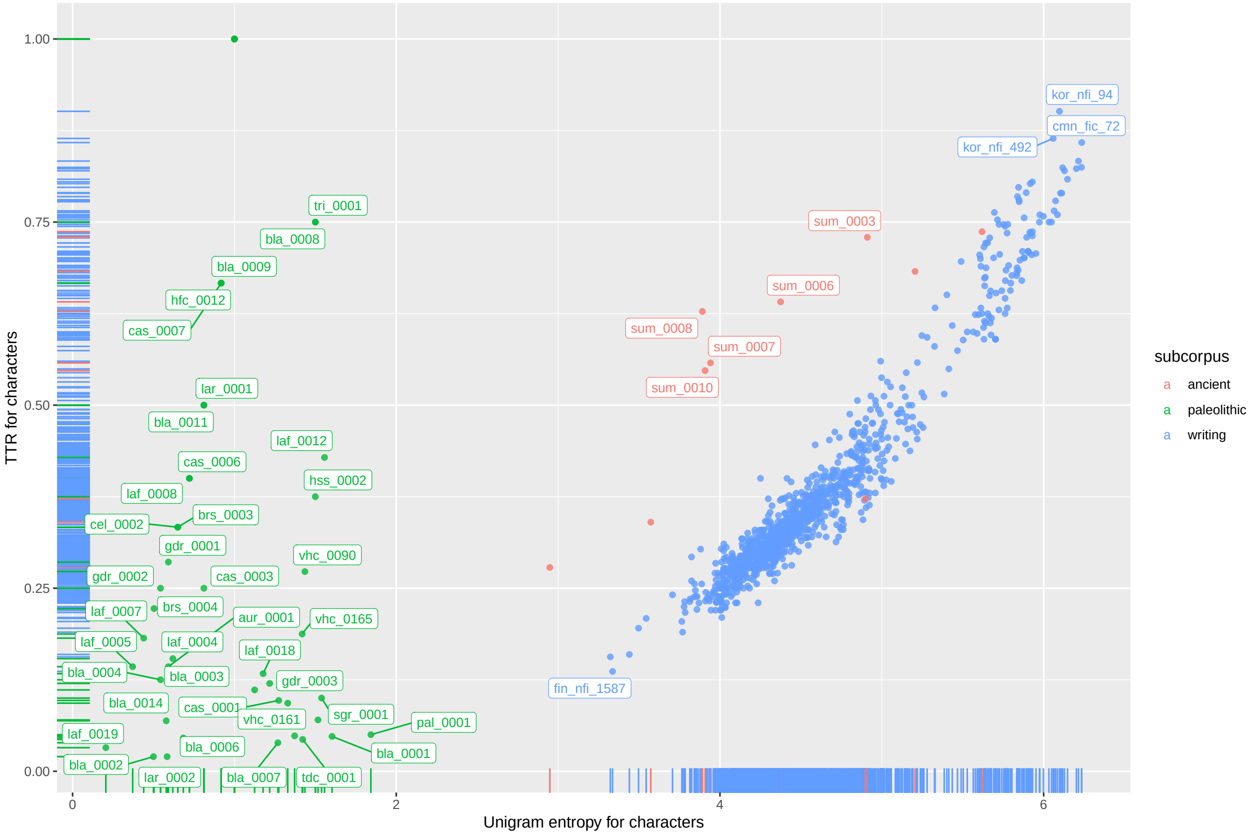
```
## Warning: Removed 7 rows containing missing values (geom_point).

## Warning: Removed 7 rows containing missing values (geom_label_repel).

## Warning: ggrepel: 1063 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

TTR vs. unigram entropy for characters

```
huni.ttr.chars.plot <- ggplot(estimations.df.combined,
                      aes(x = huni.chars, y = ttr.chars,
                      colour = subcorpus)) +
  geom_point(alpha = 0.8, size  = 1.5) +
  #theme(legend.position = "bottom") +
  geom_rug() +
  #geom_text(hjust = 0, nudge_x = 0.01, size = 2) +
  geom_label_repel(aes(label = code), force = 0.5, force.pull = 5, label.size = 0.2, size = 3) +
  labs(x = "Unigram entropy for characters", y = "TTR for characters")
huni.ttr.chars.plot
```



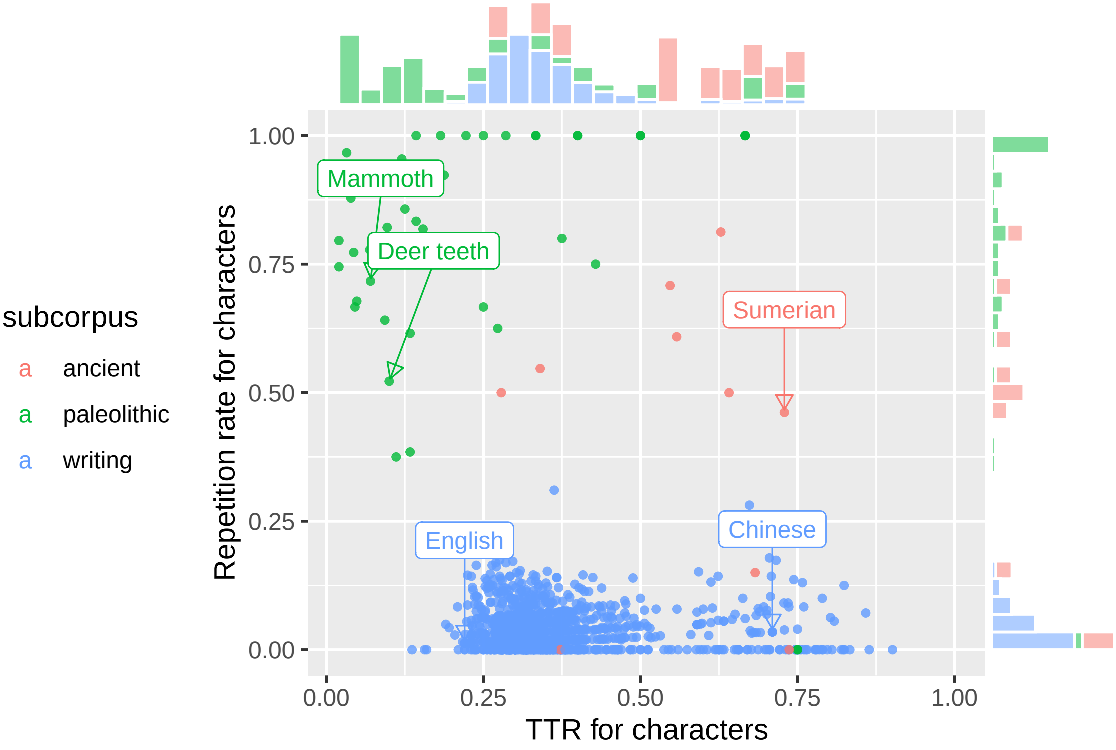
Safe complete figure to file

```
ggsave("Figures/huni_ttr_chars.pdf", huni.ttr.chars.plot, width = 13, height = 8, dpi = 300,
       scale = 1, device = cairo_pdf)
```

```
## Warning: ggrepel: 1061 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

TTR vs. repetition rate for characters

```
ttr.rm.chars.plot <- ggplot(estimations.df.combined,
                   aes(x = ttr.chars, y = rm.chars,
                   colour = subcorpus)) +
  geom_point(alpha = 0.8, size  = 1) +
  theme(legend.position = "left") +
  #geom_rug() +
  #geom_text(hjust = 0, nudge_x = 0.01, size = 2) +
  #geom_label_repel(aes(label = code), force = 0.5, force.pull = 10, label.size = 0.2, size = 3) +
  geom_label_repel(data = estimations.df.combined[estimations.df.combined$code == "sgr_0001" |
                                                  estimations.df.combined$code == "vhc_0145" |
                                                  estimations.df.combined$code == "sum_0003" |
                                                  estimations.df.combined$code == "cmn_0001" |
                                                  estimations.df.combined$code == "eng_0001" ,],
                   label = c("Deer teeth", "Mammoth", "Sumerian", "Chinese", "English"),
                   size = 3, arrow = arrow(length = unit(0.03, "npc"),
                             type = "closed", ends = "last"), nudge_y = 0.2,
                             segment.size  = 0.3) +
  labs(x = "TTR for characters", y = "Repetition rate for characters")
ttr.rm.chars.plot <- ggMarginal(ttr.rm.chars.plot, groupFill = T, type = "histogram", colour = "white")
ttr.rm.chars.plot
```

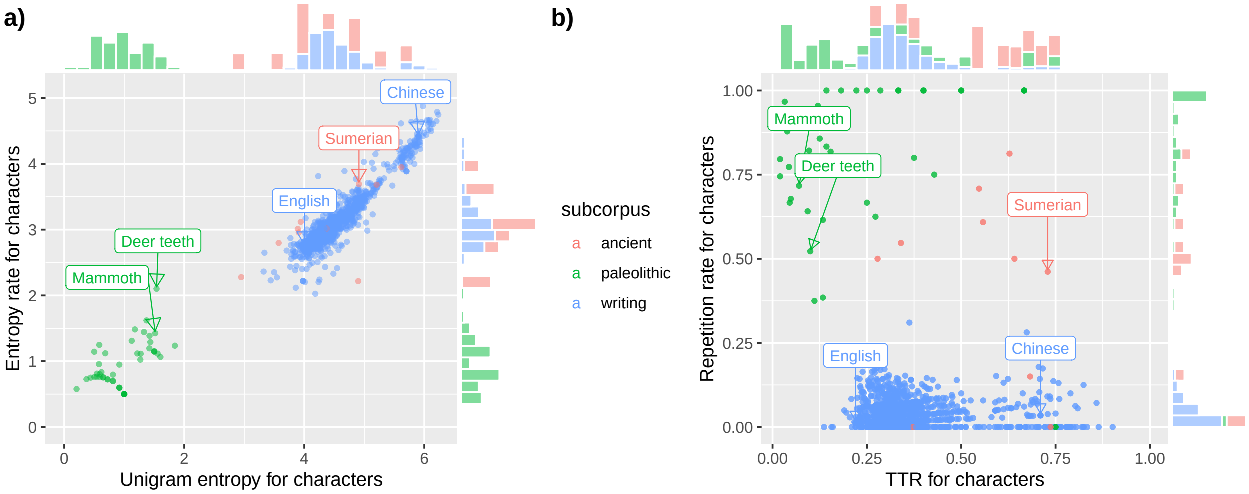


Safe complete figure to file

```
ggsave("Figures/ttr_rm_chars.pdf", ttr.rm.chars.plot, width = 6, height = 4, dpi = 300,
       scale = 1, device = cairo_pdf)
```

Combined Plots

```
plots.combined <- ggarrange(huni.hrate.chars.plot, ttr.rm.chars.plot,
                           labels = c("a)", "b)"),
                           ncol = 2, nrow = 1, widths = c(1, 1.3) )
plots.combined
```



Safe complete figure to file

```
ggsave("Figures/plots_combined.pdf", plots.combined, width = 10, height = 4, dpi = 300,
       scale = 1, device = cairo_pdf)
```