Lexical entropy quantifies discourse production severity

Kevin T Cunningham and Katarina L Haley

5/27/2019

The following analyses and visualizations were performed for Cunningham, K.T. & Haley, K.L. (2019). Lexical entropy quantifies discourse production severity. Poster presentation, *The 49th Clinical Aphasiology Conference*, Whitefish, MT, May 29, 2019. The file "data.csv" represents the data set. To replicate, one will have to read the data into R and manipulate the dataframes as you wish.

Contact: aphasia.unc.edu kevin_cunningham@med.unc.edu

RQ1: Do WIM and MATTR predict a diagnosis of aphasia?

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

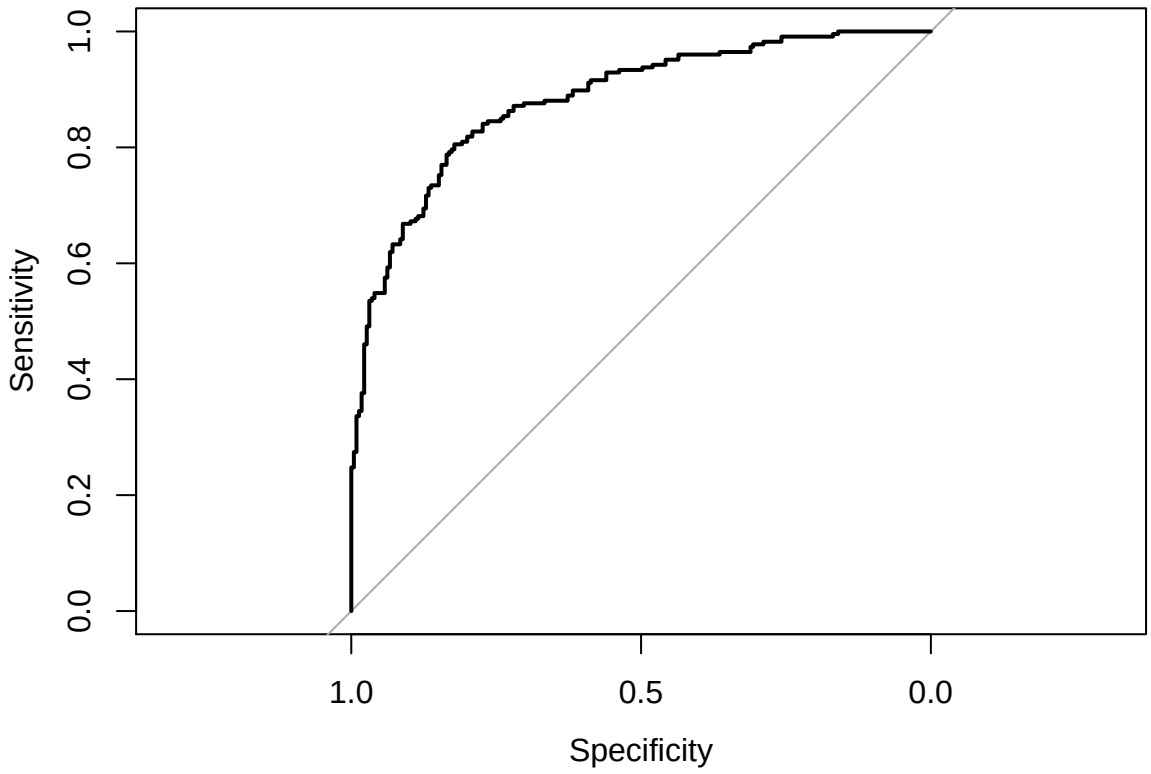
```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

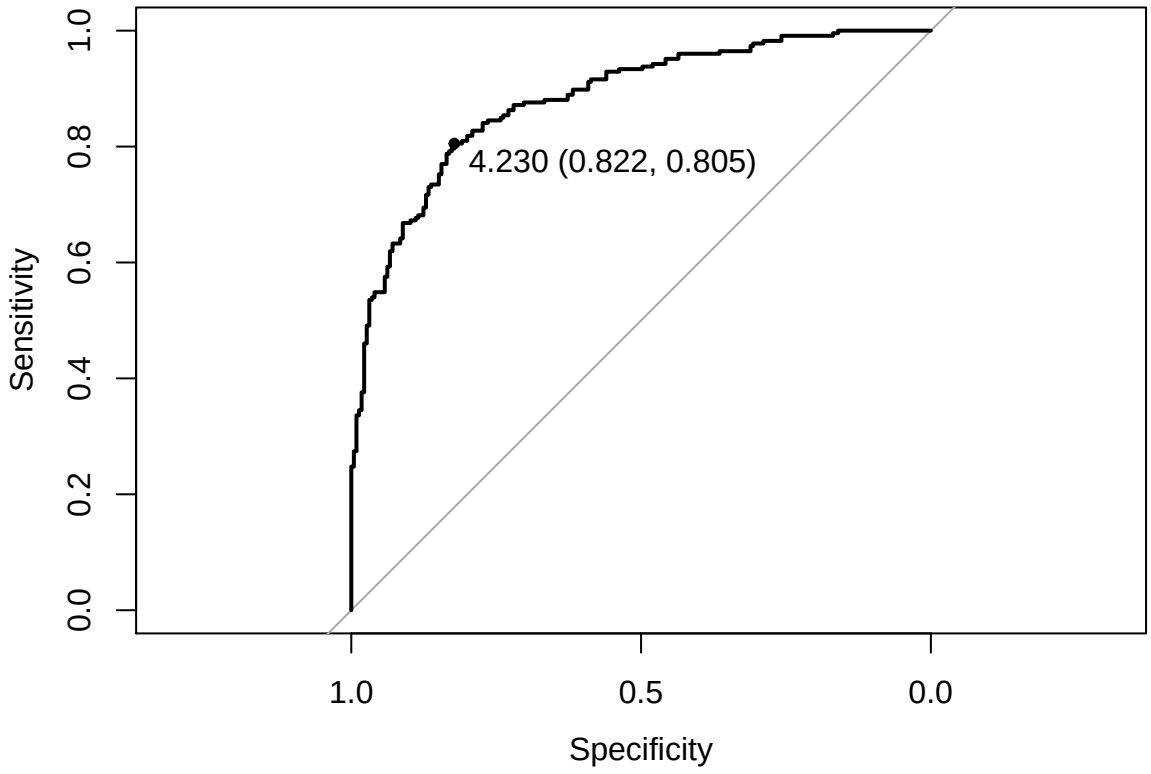
```
ent <- roc(data2$Observation, data2$ENT)
ent
```

```
##
## Call:
## roc.default(response = data2$Observation, predictor = data2$ENT)
##
## Data: data2$ENT in 225 controls (data2$Observation 0) > 226 cases (data2$Observation 1).
## Area under the curve: 0.8821
```

```
plot.roc(ent)
```



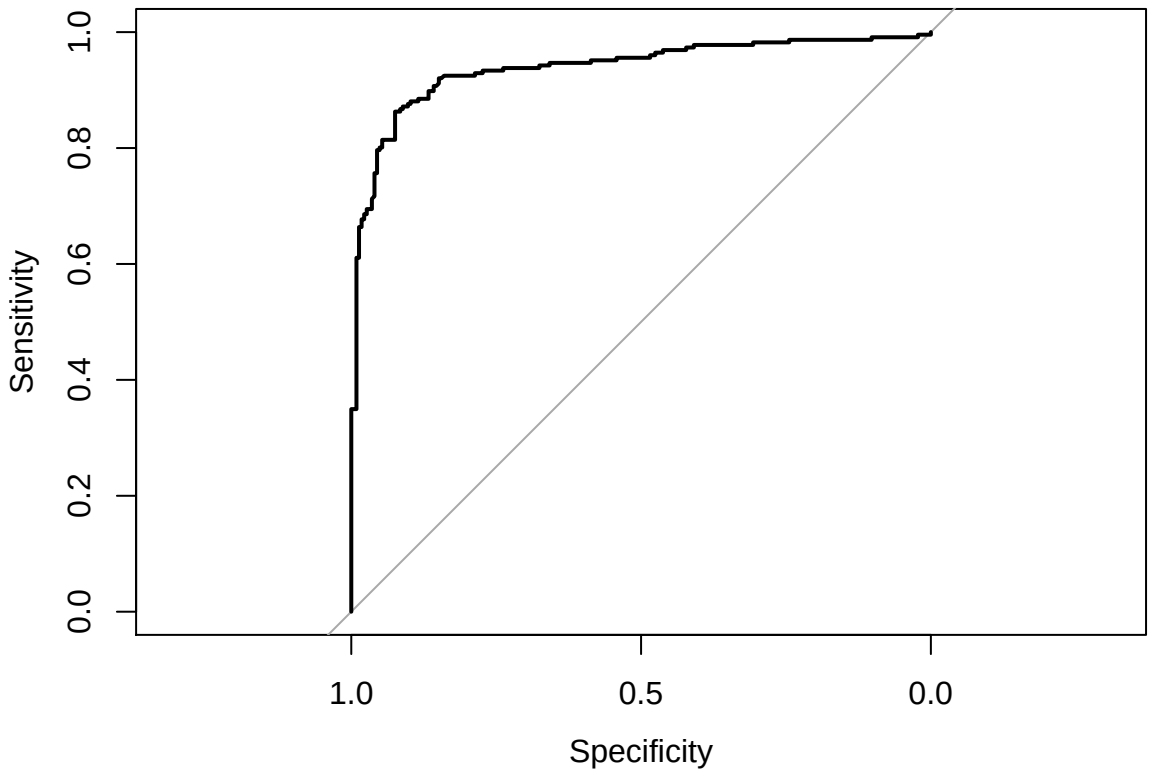

```
plot(ent, print.thres="best", print.thres.best.method="youden")
```




```
mattr <- roc(data2$Observation,data2$MATTR)
mattr
```

```
##
## Call:
## roc.default(response = data2$Observation, predictor = data2$MATTR)
##
## Data: data2$MATTR in 225 controls (data2$Observation 0) > 226 cases (data2$Observation 1).
## Area under the curve: 0.9385
```

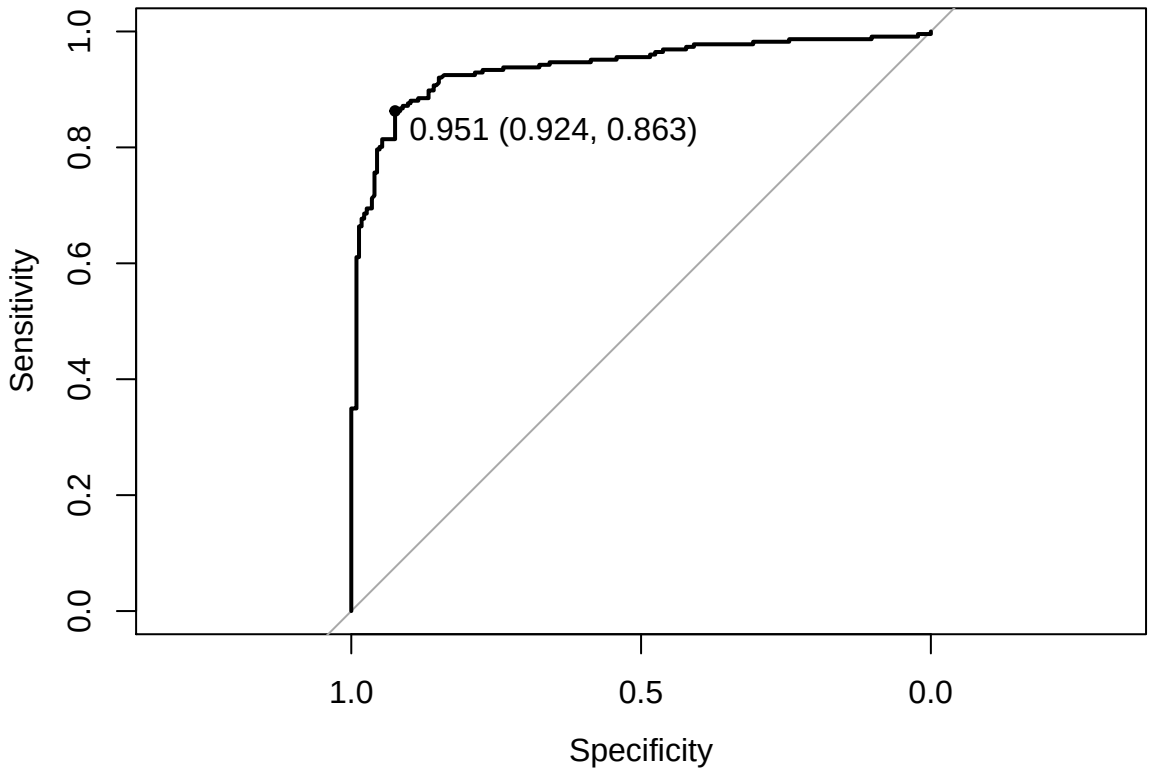
```
plot.roc(mattr)
```



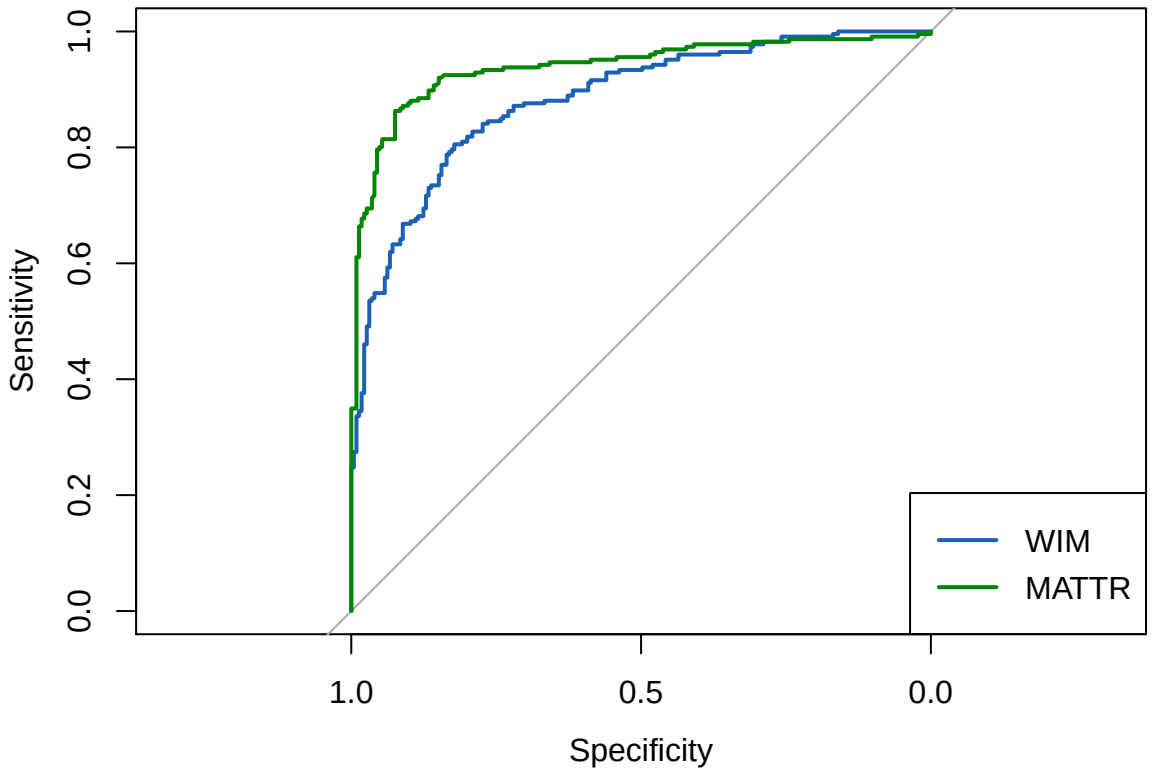

```
roc.test(ent, mattr)
```

```
##
##  DeLong's test for two correlated ROC curves
##
## data:  ent and mattr
## Z = 3.1961, p-value = 0.001393
## alternative hypothesis: true difference in AUC is not equal to 0
## sample estimates:
## AUC of roc1 AUC of roc2
##   0.8821436   0.9384562
```

```
plot(mattr, print.thres="best", print.thres.best.method="youden")
```




```
rocobj1 <- plot.roc(ent,
main="",
percent=TRUE,
col="#1c61b6")
rocobj2 <- lines.roc(mattr,
percent=TRUE,
col="#008600")
legend("bottomright", legend=c("WIM", "MATTR"), col=c("#1c61b6", "#008600", "black"), lwd=2)
```




```
#Confusion matrices for WIM based on Youden's values from ROC
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
lvs <- c("aphasia", "neurotypical")
truth <- factor(rep(lvs, times = c(226, 225)),
                levels = rev(lvs))
 pred <- factor(
     c(
         rep(lvs, times = c(181, 45)),
         rep(lvs, times = c(40, 185))),
     levels = rev(lvs))
xtab <- table(pred, truth)
WIMmatrix <-xtab

confusionMatrix(xtab, positive = "aphasia")
```

```
## Confusion Matrix and Statistics
##
##               truth
## pred           neurotypical aphasia
##   neurotypical          185      45
##   aphasia                40     181
##
##                Accuracy : 0.8115
##                  95% CI : (0.7723, 0.8466)
##     No Information Rate : 0.5011
##     P-Value [Acc > NIR] : <2e-16
```

```
##
##                  Kappa : 0.6231
##  Mcnemar's Test P-Value : 0.6644
##
##             Sensitivity : 0.8009
##             Specificity : 0.8222
##          Pos Pred Value : 0.8190
##          Neg Pred Value : 0.8043
##              Prevalence : 0.5011
##          Detection Rate : 0.4013
##    Detection Prevalence : 0.4900
##       Balanced Accuracy : 0.8116
##
##        'Positive' Class : aphasia
##
```

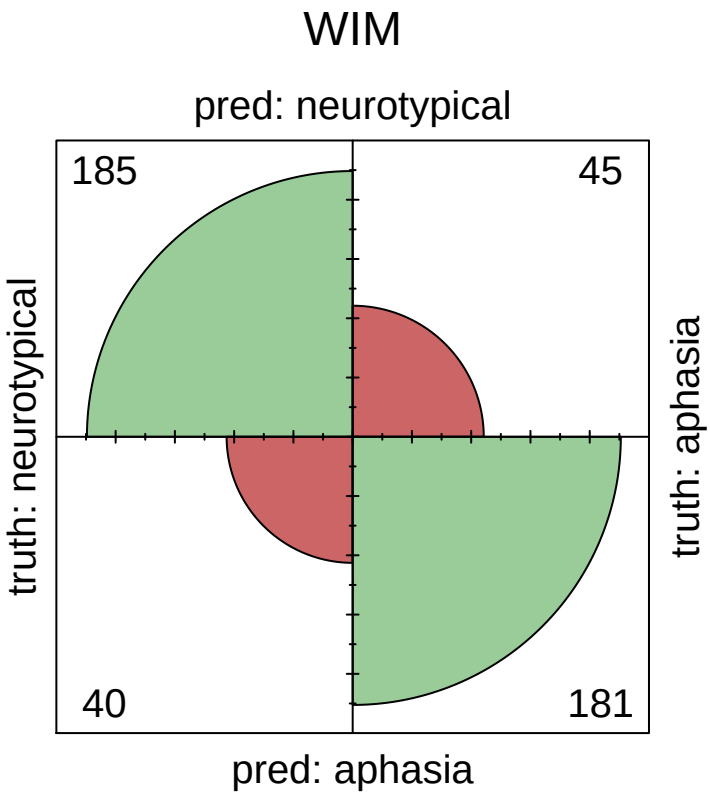
```
#Confusion matrices for MATTR based on Youden's values from ROC
library(caret)
lvs <- c("aphasia", "neurotypical")
truth <- factor(rep(lvs, times = c(226, 225)),
                levels = rev(lvs))
 pred <- factor(
     c(
         rep(lvs, times = c(194, 32)),
         rep(lvs, times = c(19, 206))),
     levels = rev(lvs))
xtab <- table(pred, truth)
MATTRmatrix <- xtab
confusionMatrix(xtab, positive = "aphasia")
```

```
## Confusion Matrix and Statistics
##
##               truth
## pred           neurotypical aphasia
##   neurotypical          206      32
##   aphasia                19     194
##
##                Accuracy : 0.8869
##                  95% CI : (0.854, 0.9146)
##     No Information Rate : 0.5011
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.7739
##  Mcnemar's Test P-Value : 0.09289
##
##             Sensitivity : 0.8584
##             Specificity : 0.9156
##          Pos Pred Value : 0.9108
##          Neg Pred Value : 0.8655
##              Prevalence : 0.5011
##          Detection Rate : 0.4302
##    Detection Prevalence : 0.4723
##       Balanced Accuracy : 0.8870
##
##        'Positive' Class : aphasia
```

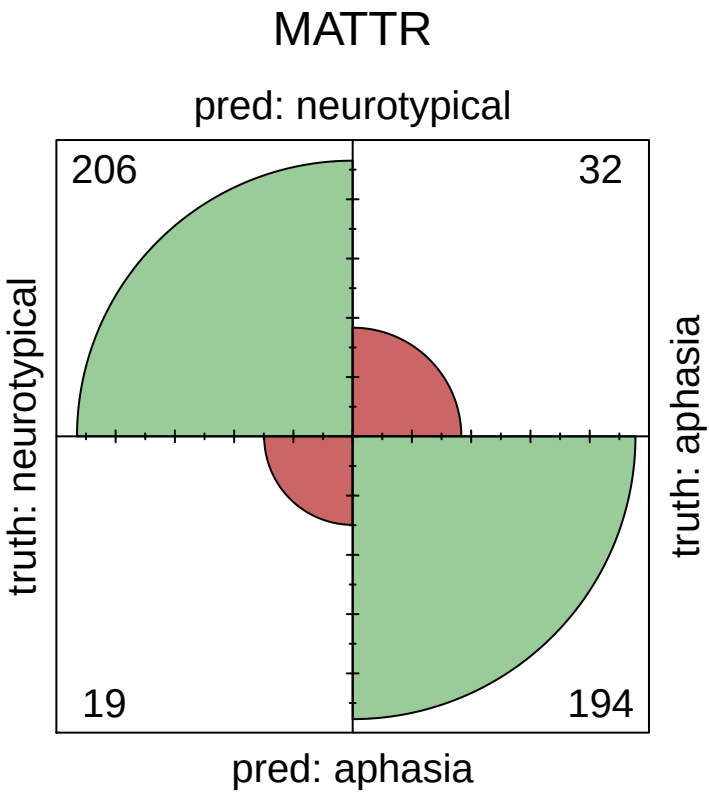
##

visualization of confusion matrices

```
fourfoldplot(WIMmatrix, color = c("#CC6666", "#99CC99"),conf.level = 0, margin = 1, main = "WIM")
```




```
fourfoldplot(MATTRmatrix, color = c("#CC6666", "#99CC99"),conf.level = 0, margin = 1, main = "MATTR")
```



RQ2: Are WIM and MATTR sensitive to expected variation of discourse deficits among people with aphasia?

data cleaning note: will need to remove bu11a before these analyses. no wab.

ANOVA of WIM on WAB Subtype. Calculate Eta-squared effect size.

```
data3 <- read.csv("ENTAphasia.csv")
mod <- lm (data3$ENT~data3$WABType)
analysis <- anova(mod)
pairwise.t.test(data3$ENT, data3$WABType, p.adj= "holm")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  data3$ENT and data3$WABType
##
##            Anomic  Broca   Conduction
## Broca      9.1e-10 -       -
## Conduction 1       1.9e-08 -
## Wernicke   1       8.1e-06 1
##
## P value adjustment method: holm
```

```
#Effect size
library(sjstats)
```

```
## Warning: package 'sjstats' was built under R version 3.5.2
```

```
## Warning in checkMatrixPackageVersion(): Package version inconsistency detected.
## TMB was built with Matrix version 1.2.15
## Current Matrix version is 1.2.14
## Please re-install 'TMB' from source using install.packages('TMB', type = 'source') or ask CRAN for a
```

```
omega_sq(mod)
```

```
##            term omegasq
## 1 data3$WABType   0.198
```

```
mod <- lm (data3$MATTR~data3$WABType)
analysis <- anova(mod)
pairwise.t.test(data3$MATTR, data3$WABType, p.adj= "holm")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  data3$MATTR and data3$WABType
##
##            Anomic Broca Conduction
## Broca      0.1    -     -
## Conduction 1.0    1.0   -
## Wernicke   1.0    1.0   1.0
##
## P value adjustment method: holm
```

```
#Effect size
library(sjstats)
omega_sq(mod)
```

```
##            term omegasq
## 1 data3$WABType   0.013
```

ANOVA of WIM on WAB Severity

```
mod <- lm(data4$ENT~data4$Severity)
analysis <- anova(mod)
analysis
```

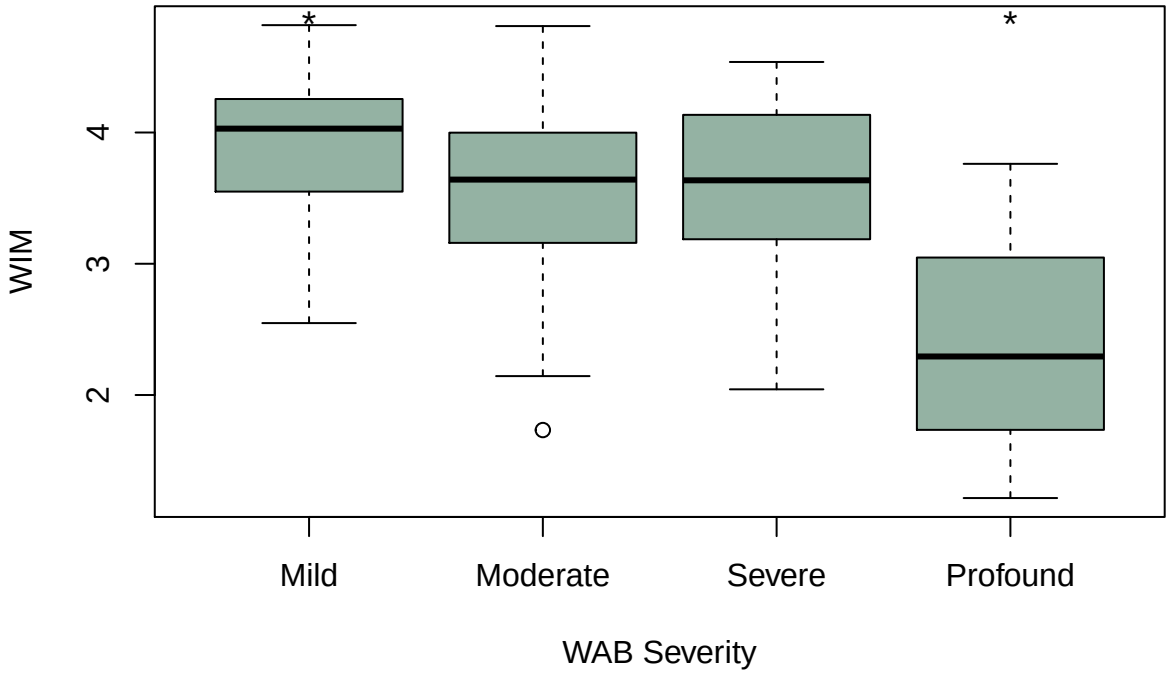
```
## Analysis of Variance Table
##
## Response: data4$ENT
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## data4$Severity   3 13.862  4.6206  12.355 1.677e-07 ***
## Residuals      221 82.649  0.3740
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pairwise.t.test(data4$ENT, data4$Severity, p.adj= "holm")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  data4$ENT and data4$Severity
##
##          Mild    Moderate Profound
## Moderate 0.00064 -        -
```

```
## Profound 1.4e-05 0.00094  -
## Severe   0.00472 0.62381  0.00272
##
## P value adjustment method: holm
```

```
data4$Severity <- factor(data4$Severity, c("Mild", "Moderate", "Severe", "Profound"))
boxplot(data4$ENT~data4$Severity, col=rgb(0.3,0.5,0.4,0.6), ylab="WIM",xlab="WAB Severity")
mtext("*", side=3, line=-1, at=1, cex=1.2)
mtext("*", side=3, line=-1, at=4, cex=1.2)
```




```
#Effect size
library(sjstats)
omega_sq(mod)
```

```
##             term omegasq
## 1 data4$Severity   0.131
```

ANOVA of MATTR on WAB Severity

```
mod <- lm(data4$MATTR~data4$Severity)
mod
```

```
##
## Call:
## lm(formula = data4$MATTR ~ data4$Severity)
##
## Coefficients:
##             (Intercept)  data4$SeverityModerate    data4$SeveritySevere
##                 0.90946                -0.01568                -0.02004
## data4$SeverityProfound
##               -0.07443
```

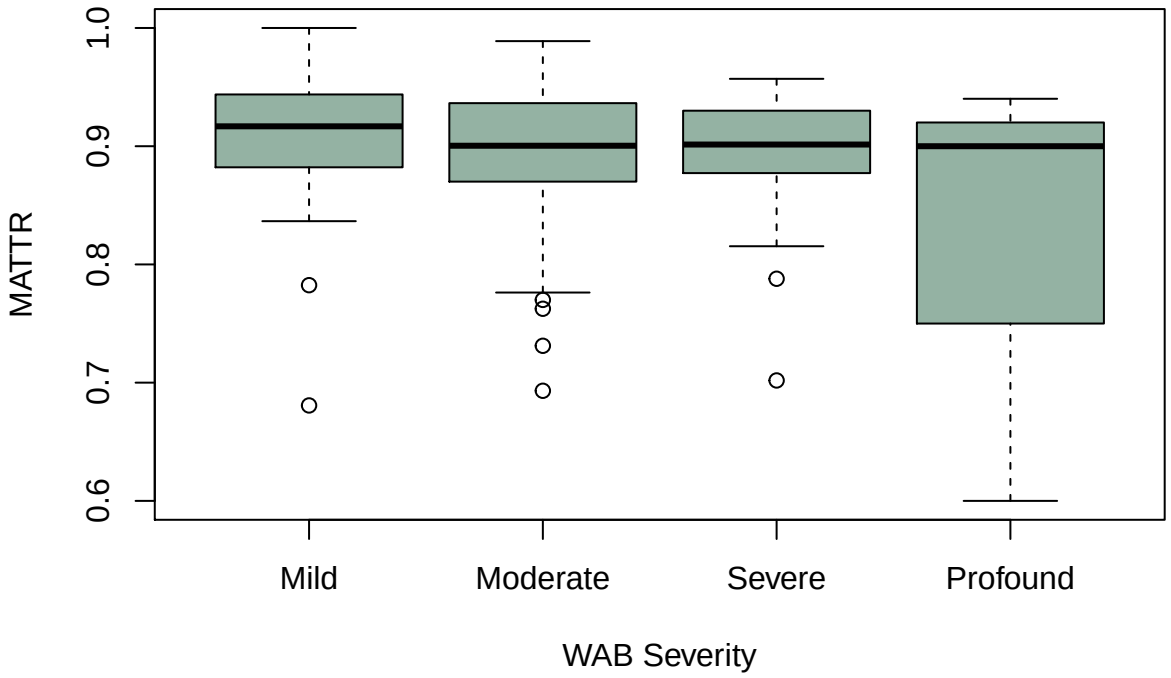
```
analysis <- anova(mod)
analysis
```

```
## Analysis of Variance Table
##
## Response: data4$MATTR
##                  Df  Sum Sq   Mean Sq F value  Pr(>F)
## data4$Severity    3 0.03234 0.0107798  3.4296 0.01791 *
## Residuals       221 0.69464 0.0031432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pairwise.t.test(data4$MATTR, data4$Severity, p.adj= "holm")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  data4$MATTR and data4$Severity
##
##          Mild  Moderate Severe
## Moderate 0.211 -        -
## Severe   0.215 0.722    -
## Profound 0.059 0.206    0.215
##
## P value adjustment method: holm
```

```
data4$Severity <- factor(data4$Severity, c("Mild", "Moderate", "Severe", "Profound"))
boxplot(data4$MATTR~data4$Severity, col=rgb(0.3,0.5,0.4,0.6), ylab="MATTR",xlab="WAB Severity")
```



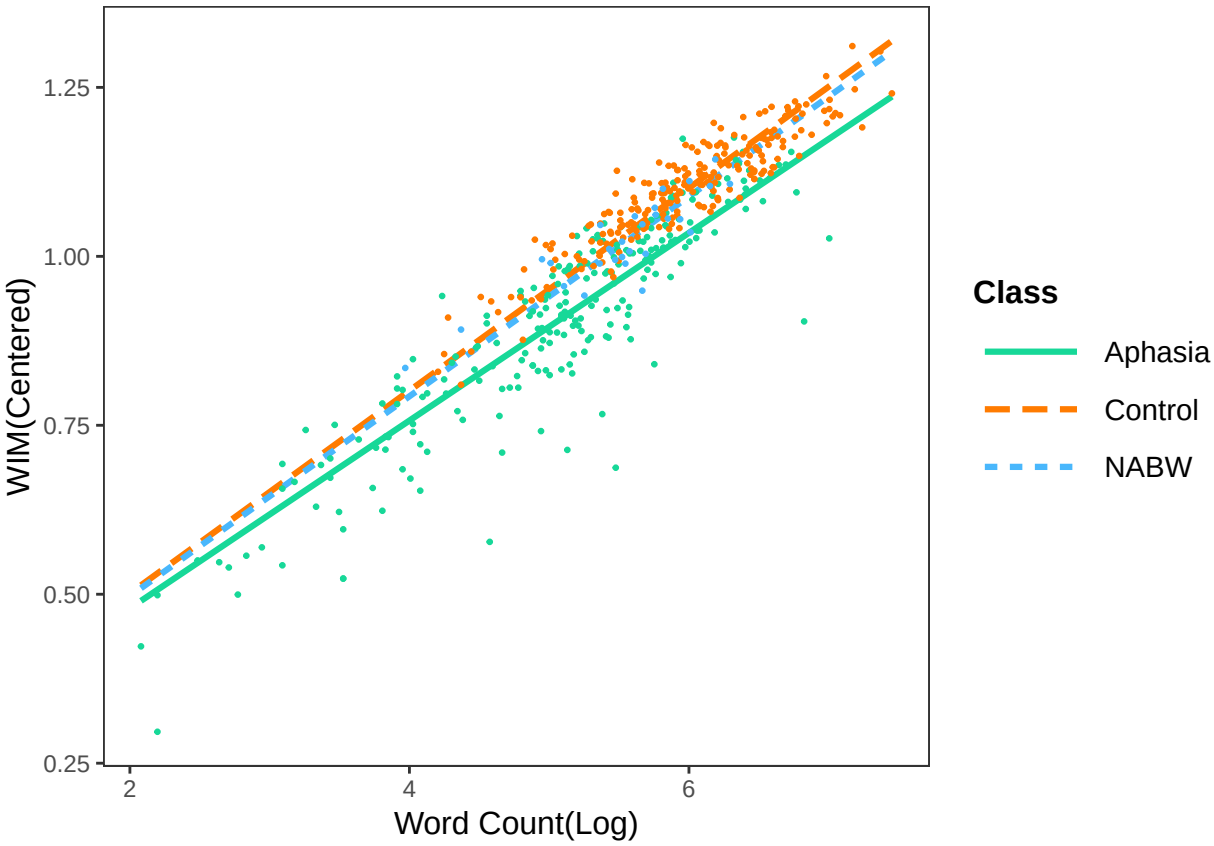

```
#Effect size
library(sjstats)
omega_sq(mod)
```

```
##             term omegasq
## 1 data4$Severity   0.031
```

RQ3: Do length effects on WIM and MATTR vary among PWA, NABW, and NORM?

```
logword <- log(data$WORD)
datainteract <- cbind(logword, data)

model5.c<-lm(ENT_C~logword:Class, data=datainteract)
interact_plot(model5.c,pred = logword, modx = Class, modx.values = c("Aphasia", "Control", "NABW"), y.la
```




```
model6.c <- lm(MATTR_C~WORD:Class, data=data)
summary(model6.c)
```

```
##
## Call:
## lm(formula = MATTR_C ~ WORD:Class, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34431 -0.01813  0.00964  0.03118  0.08435
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         9.867e-01  3.915e-03 252.043  < 2e-16 ***
## WORD:ClassAphasia -4.746e-05  1.503e-05  -3.157  0.00169 **
## WORD:ClassControl  7.815e-05  8.527e-06   9.165  < 2e-16 ***
## WORD:ClassNABW     7.434e-05  3.292e-05   2.258  0.02438 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.04908 on 475 degrees of freedom
## Multiple R-squared:  0.2527, Adjusted R-squared:  0.248
## F-statistic: 53.55 on 3 and 475 DF,  p-value: < 2.2e-16
```

```
interact_plot(model6.c,pred = WORD, modx = Class, modx.values = c("Aphasia", "Control", "NABW"), y.label
```