Eye-movement-analysis-example-final

April 4, 2017

1 Example of using R Markdown with jupyter notebook: An eye movement analysis of sentence reading, comparing a reader with aphasia to a neurologically healthy reader

1.1 Abstract

Mild reading difficulties are a pervasive symptom of aphasia, a language impairment common post stroke. In this study, we used eye tracking to investigate sentence reading by one person diagnosed with aphasia (PWA), compared to a neurologically healthy participant (NHI). Data were extracted from a larger project on sentence reading (published in Aphasiology online). The main aim of this study was to find out whether the eye movements of these two readers are influenced by linguistic factors of word frequency and contextual predictability. The two participants read sentences including target words that varied in word frequency and contextual predictability, and answered comprehension questions. We recorded gaze duration, total fixation duration, and first-pass regressions. Results demonstrated that the PWA had prolonged gaze and total fixation durations and an increase of first-pass regressions compared to the NHI. Both readers were influenced by word frequency and predictability, but in different ways. Readers varied in gaze duration and first-pass regressions in particular, which may point to differences in the phase of lexical access.

1.2 Load libraries

```
In [111]: library(gdata) # used
          library(psych)
          library(compute.es)
          library(ggplot2) # used
          library(multcomp)
          library(pastecs) # used
          library(ez)
          library(Hmisc)
          library(reshape) # used
          library(gridExtra) #used
          library(lme4) #used
          library(lmerTest) #used
```

```
In [112]: #set working directory
          setwd("~/code/juypternotebooks")
```

1.3 We are going to load data of the two participants from the reading study.

1.3.1 Open database:

```
In [113]: rawdata=read.xls("EMdataexample.xlsx",
                           na.strings = c("zero"),
                           colClasses = c(
                             'factor',   # RECORDING_SESSION_LABEL
                             'factor',   # GROUP
                             'factor',   # ID_OVERALL
                             'factor',   # ID
                             'factor',   # TRIAL_INDEX
                             'factor',   # trial_type
                             'factor',   # FREQUENCY
                             'factor',   # PREDICTABILITY
                             'factor',   # SENTENCE
                             'factor',   # ITEM
                             'factor',   # QUESTION
                             'factor',   # CRITICAL_WORD
                             'factor',   # ACCURACY
                             'character', # SINGLE_FIXATION_DURATION
                             'character', # FIRST_FIXATION_DURATION
                             'character', # GAZE_DURATION
                             'character', # RIGHT_BOUNDED_DURATION
                             'character', # REGRESSION_PATH_DURATION
                             'character', # REREADING_DURATION
                             'character', # TOTAL_DURATION
                             'character',  # FIRST_PASS_REGRESSION
                             'character',  # FIRST_PASS_FIXATION
                             'factor',   # FIRST_PASS_MULTI_FIXATION
                             'character' #trials.fixated
                           )
          )
```

```
In [114]: # rawdata
```

```
In [115]: ## Create a new dataframe for analysis
```

```
In [116]: data <-rawdata
```

1.4 Explore the data

```
In [117]: #str(data)
          #summary(data)
          #head(data[, 1:10])
          #tail(data[, 1:10])
          #dim(data)
```

1.5 Preparing variables we are interested in:

1.5.1 Create variables as numeric

```
In [118]: data$TOTAL_DURATION  <-as.numeric(data$TOTAL_DURATION)
          data$GAZE_DURATION  <-as.numeric(data$GAZE_DURATION)
          data$FIRST_PASS_REGRESSION <-as.numeric(data$FIRST_PASS_REGRESSION)
```

```
Warning message in eval(expr, envir, enclos):
"NAs introduced by coercion"
```

1.5.2 Check whether the data frame inlcudes NAs

```
In [119]: #which(is.na(data$GAZE_DURATION))
          #which(is.na(data$TOTAL_DURATION))
          #which(is.na(data$FIRST_PASS_REGRESSION))
          #which(is.na(data$FIRST_PAST_FIXATION))
```

1.5.3 Exclude the NAs

```
In [120]: data <- data[(!is.na(data$GAZE_DURATION)),]
          data <- data[(!is.na(data$TOTAL_DURATION)),]
          data <- data[(!is.na(data$FIRST_PASS_REGRESSION)),]
          data <- data[(!is.na(data$FIRST_PAST_FIXATION)),]
```

1.5.4 Check whether it worked ok

```
In [121]: ##which(is.na(data$GAZE_DURATION))
          ##which(is.na(data$TOTAL_DURATION))
          ##which(is.na(data$FIRST_PASS_REGRESSION))
          ##which(is.na(data$FIRST_PAST_FIXATION))
```

1.5.5 Rename GROUP as CASE - because this example dataset is restricted to the comparison of two cases

```
In [122]: data <- rename(data, c(GROUP="CASE"))
```

```
In [123]: # data
```

1.6 Data analysis

There are four conditions (=TRIAL TYPES) in this dataset. Sentences with: * High frequency predictable words * High frequency unpredictable words * Low frequency predictable words * Low frequency unpredictable words

Independent variables are: word frequency, contextual predictability and case

Dependent variables are: gaze duration, total fixation duration, first-pass regression

1.6.1 We start by getting some descriptive stats, comparing the four trial types:

1.6.2 Gaze duration as a measure of TRIAL TYPE and CASE

```
In [124]: # by(data$GAZE_DURATION, list(data$TRIAL_TYPE, data$CASE), stat.desc, bas
```

1.6.3 Total fixation duration as a measure of TRIAL TYPE and CASE

```
In [125]: # by(data$TOTAL_DURATION, list(data$TRIAL_TYPE, data$CASE), stat.desc, ba
```

1.6.4 First-pass regression as a measure of TRIAL TYPE and CASE

```
In [126]: # by(data$FIRST_PASS_REGRESSION, list(data$TRIAL_TYPE, data$CASE), stat.d
```

1.7 Plotting - Eye movements independent of trial types

1.7.1 Gaze duration

```
In [127]: plot_gaze <- ggplot(data,aes(x=CASE, y=GAZE_DURATION, fill=CASE)) +
            stat_summary(fun.data=mean_cl_normal,position=position_dodge(0.95),geom
            stat_summary(fun.y=mean,position=position_dodge(width=0.95),geom="bar")
              ylab("Gaze duration in ms") +
              xlab("Case") +
              theme(axis.text=element_text(size=13)) +
              theme(axis.title.x=element_text(size=13)) +
              theme(axis.title.y=element_text(size=13)) +
              ggtitle("Gaze duration NHI vs PWA")
          #plot_gaze
```

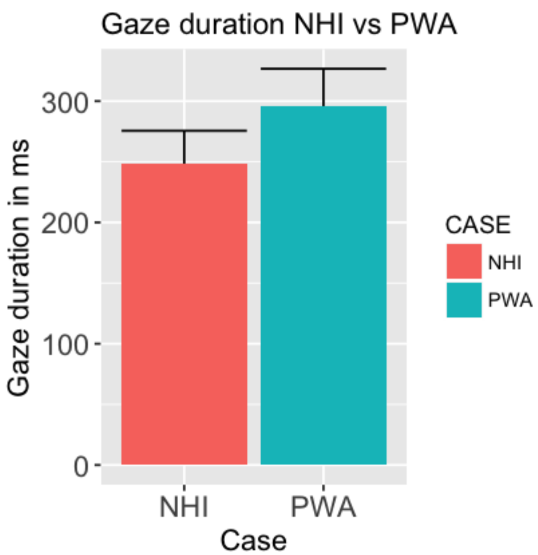
1.7.2 Total fixation duration

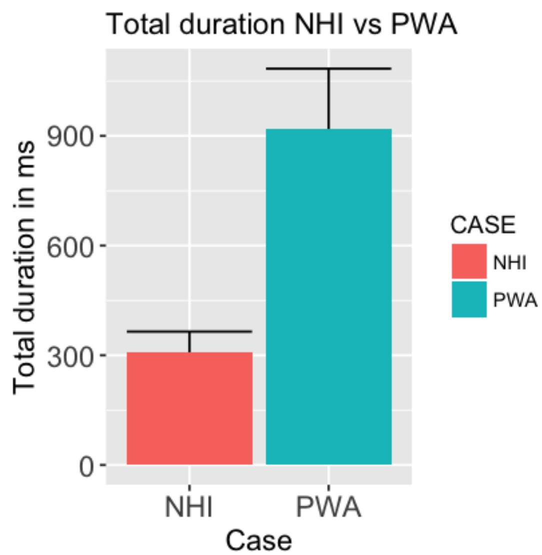
```
In [128]: plot_total <- ggplot(data,aes(x=CASE, y=TOTAL_DURATION, fill=CASE)) +
            stat_summary(fun.data=mean_cl_normal,position=position_dodge(0.95),geom
            stat_summary(fun.y=mean,position=position_dodge(width=0.95),geom="bar")
              ylab("Total duration in ms") +
              xlab("Case") +
              theme(axis.text=element_text(size=13)) +
              theme(axis.title.x=element_text(size=13)) +
              theme(axis.title.y=element_text(size=13)) +
              ggtitle("Total duration NHI vs PWA")
          #plot_total
```

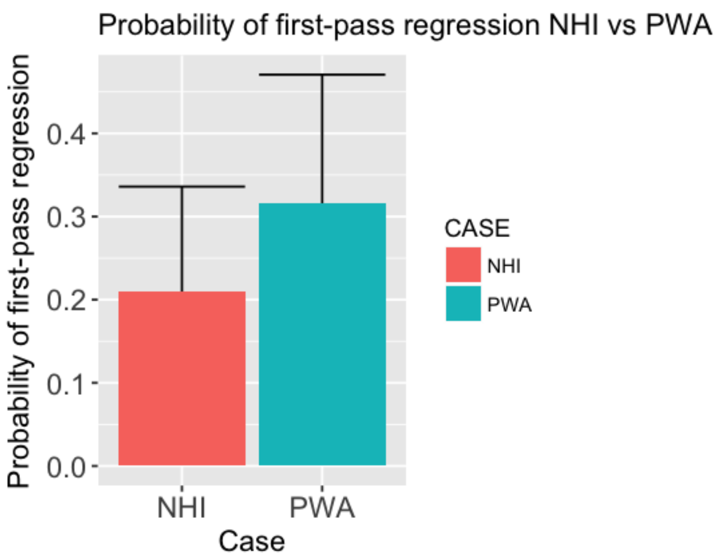
1.7.3 First-pass regression

```
In [129]: plot_regress_prob <- ggplot(data,aes(x=CASE, y=FIRST_PASS_REGRESSION, fil
            stat_summary(fun.data=mean_cl_normal,position=position_dodge(0.95),geom
            stat_summary(fun.y=mean,position=position_dodge(width=0.95),geom="bar")
              ylab("Probability of first-pass regression") +
              xlab("Case") +
              theme(axis.text=element_text(size=13)) +
              theme(axis.title.x=element_text(size=13)) +
              theme(axis.title.y=element_text(size=13)) +
              ggtitle("Probability of first-pass regression NHI vs PWA")
          #plot_regress_prob
```

```
In [130]: grid.arrange(plot_gaze, plot_total, plot_regress_prob, ncol=2, respect=TR
```







1.7.4 Summary:

The participant with aphasia shows an increase in reading times and in first-pass regressions.

1.8 Plotting - Eye movements as a function of trial type

1.8.1 Gaze duration

```
In [131]: line_gaze <-
          ggplot(data, aes(x=PREDICTABILITY, y=GAZE_DURATION, group=interaction(CAS
            stat_summary(fun.data=mean_cl_normal, geom="errorbar", position=positic
            stat_summary(fun.y=mean, geom="line", position=position_dodge(width=0.1
            stat_summary(fun.y=mean, geom="point",position=position_dodge(width=0.1
```

```
  scale_shape_manual(values = c(16, 18)) +
  scale_x_discrete(limits=c("predictable", "unpredictable")) +
  theme (axis.text.x=element_text(colour="#000000", size=13)) +
  theme (axis.text.y=element_text(colour="#000000", size=13)) +
  theme(axis.title.y=element_text(colour="#000000", size=13)) +
  theme (axis.title.x = element_blank()) +
  scale_y_continuous(name="Gaze duration in ms") +
  theme(legend.title = element_text(size=13)) +
  theme(legend.text = element_text(size = 13)) +
  theme(legend.position="right")
#line_gaze
```

1.8.2 Total fixation duration

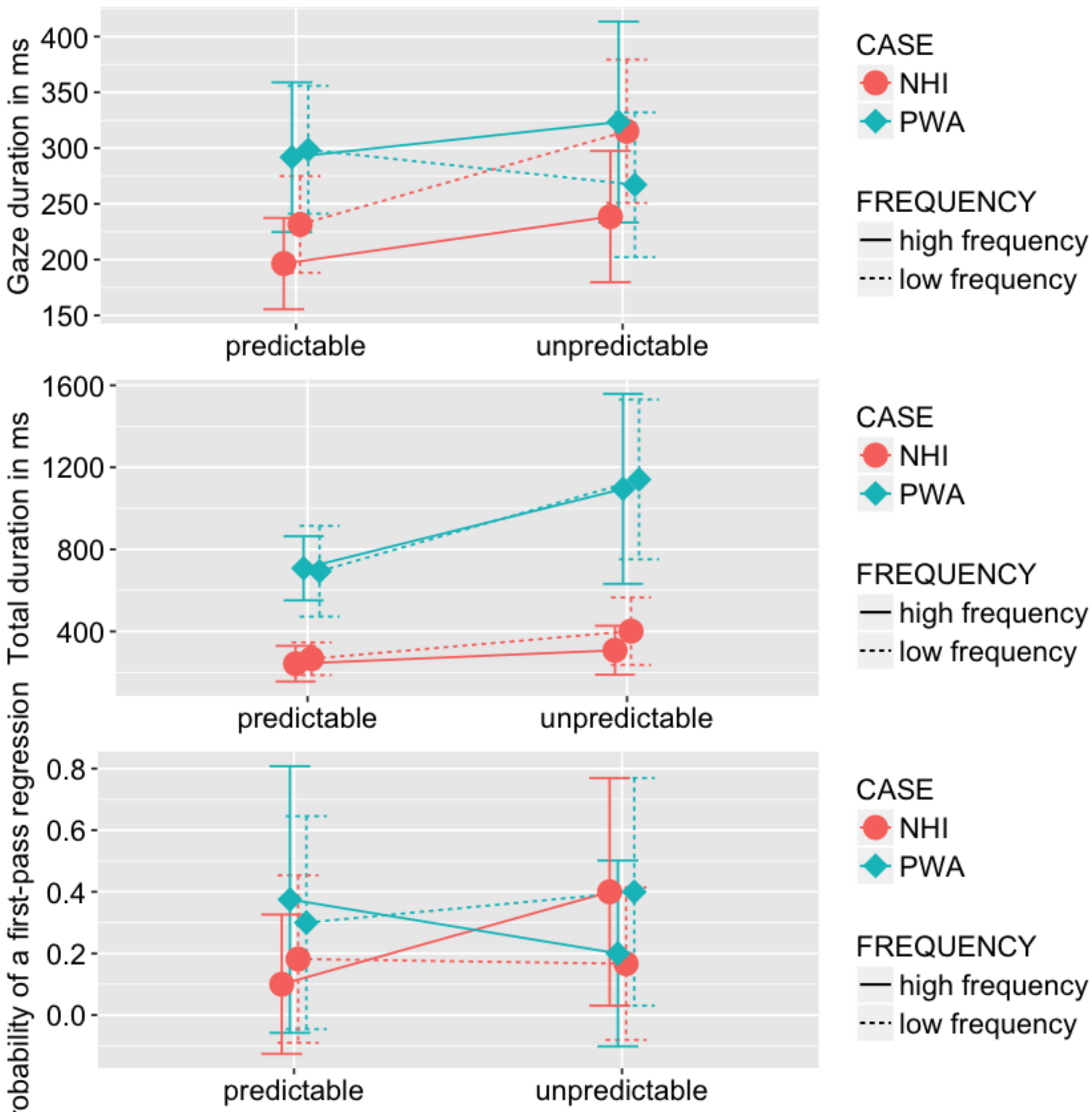
```
In [132]: line_total <-
          ggplot(data, aes(x=PREDICTABILITY, y=TOTAL_DURATION, group=interaction(CA
            stat_summary(fun.data=mean_cl_normal, geom="errorbar", position=positic
            stat_summary(fun.y=mean, geom="line", position=position_dodge(width=0.1
            stat_summary(fun.y=mean, geom="point",position=position_dodge(width=0.1
            scale_shape_manual(values = c(16, 18)) +
            scale_x_discrete(limits=c("predictable", "unpredictable")) +
            theme (axis.text.x=element_text(colour="#000000", size=13)) +
            theme (axis.text.y=element_text(colour="#000000", size=13)) +
            theme(axis.title.y=element_text(colour="#000000", size=13)) +
            theme (axis.title.x = element_blank()) +
            scale_y_continuous(name="Total duration in ms") +
            theme(legend.title = element_text(size=13)) +
            theme(legend.text = element_text(size = 13)) +
            theme(legend.position="right")
          #line_total
```

1.8.3 First-pass regression

```
In [133]: line_regression <-
          ggplot(data, aes(x=PREDICTABILITY, y=FIRST_PASS_REGRESSION, group=interac
            stat_summary(fun.data=mean_cl_normal, geom="errorbar", position=positic
            stat_summary(fun.y=mean, geom="line", position=position_dodge(width=0.1
            stat_summary(fun.y=mean, geom="point",position=position_dodge(width=0.1
            scale_shape_manual(values = c(16, 18)) +
            scale_x_discrete(limits=c("predictable", "unpredictable")) +
            theme (axis.text.x=element_text(colour="#000000", size=13)) +
            theme (axis.text.y=element_text(colour="#000000", size=13)) +
            theme(axis.title.y=element_text(colour="#000000", size=13)) +
            theme (axis.title.x = element_blank()) +
            scale_y_continuous(name="Probability of a first-pass regression") +
            theme(legend.title = element_text(size=13)) +
            theme(legend.text = element_text(size = 13)) +
            theme(legend.position="right")
```

```
#line_regression
```

```
In [134]: grid.arrange(line_gaze, line_total, line_regression, nrow=3)
```



1.8.4 Linear mixed model analysis of effects of word frequency and predictability

1.8.5 Gaze duration

```
In [135]: model_simple = lmer (GAZE_DURATION ~ CASE + (1 | ITEM),
                                    data=data, REML=FALSE)
          ##summary(model_simple)
```

```
In [136]: model_a = lmer (GAZE_DURATION ~CASE+FREQUENCY + (1 | ITEM),
                                    data=data, REML=FALSE)
```

```
In [137]: anova(model_simple, model_a)
          # not significant so FREQUENCY does not improve model fit
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
object	4	967.9236	977.5014	-479.9618	959.9236	NA	NA	NA
..1	5	969.1437	981.1159	-479.5718	959.1437	0.7799613	1	0.377153

```
In [138]: model_b = lmer (GAZE_DURATION ~CASE+PREDICTABILITY + (1 | ITEM),
                                      data=data, REML=FALSE)

In [139]: anova(model_simple, model_b)
          # not significant so PREDICTABILITY does not improve model fit
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
object	4	967.9236	977.5014	-479.9618	959.9236	NA	NA	NA
..1	5	966.7371	978.7093	-478.3685	956.7371	3.18653	1	0.07424748

```
In [140]: # Checking for interaction between CASE and PREDICTABILITY
          model_c = lmer (GAZE_DURATION~CASE + PREDICTABILITY + (1 | ITEM),
                                    data=data, REML=FALSE)
          model_d = lmer (GAZE_DURATION~CASE * PREDICTABILITY + (1 | ITEM),
                                    data=data, REML=FALSE)
          anova(model_c, model_d) # not significant so no interaction between CASE
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
object	5	966.7371	978.7093	-478.3685	956.7371	NA	NA	NA
..1	6	965.8396	980.2063	-476.9198	953.8396	2.897467	1	0.08871884

```
In [141]: # Checking for interaction between CASE and FREQUENCY
          model_e = lmer (GAZE_DURATION~CASE + FREQUENCY + (1 | ITEM),
                                    data=data, REML=FALSE)
          model_f = lmer (GAZE_DURATION~CASE * FREQUENCY + (1 | ITEM),
                                    data=data, REML=FALSE)
          anova(model_e, model_f) # there is a significant interaction between CASI
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
object	5	969.1437	981.1159	-479.5718	959.1437	NA	NA	NA
..1	6	966.6410	981.0077	-477.3205	954.6410	4.502692	1	0.03384154

1.8.6 Total duration

```
In [142]: model_TD_simple = lmer (TOTAL_DURATION ~CASE + (1 | ITEM),
                                      data=data, REML=FALSE)
          # summary(model_TD_simple)

In [143]: model_TD_a = lmer (TOTAL_DURATION ~CASE+FREQUENCY + (1 | ITEM),
                                      data=data, REML=FALSE)

In [144]: anova(model_TD_simple, model_TD_a)
          # not significant so FREQUENCY does not improve model fit
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
object	4	1192.489	1202.067	-592.2445	1184.489	NA	NA	NA
..1	5	1194.352	1206.324	-592.1758	1184.352	0.1373259	1	0.710954

```
In [145]: model_TD_b = lmer (TOTAL_DURATION ~CASE+PREDICTABILITY + (1 | ITEM),
                                 data=data, REML=FALSE)
```

```
In [146]: anova(model_TD_simple, model_TD_b)
          # significant so PREDICTABILITY does improve model fit
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
object	4	1192.489	1202.067	-592.2445	1184.489	NA	NA	NA
..1	5	1184.056	1196.028	-587.0281	1174.056	10.43276	1	0.001237993

```
In [147]: # Checking for interaction between CASE and PREDICTABILITY
          model_TD_c = lmer (TOTAL_DURATION~CASE + PREDICTABILITY + (1 | ITEM),
                               data=data, REML=FALSE)
          model_TD_d = lmer (TOTAL_DURATION~CASE * PREDICTABILITY + (1 | ITEM),
                               data=data, REML=FALSE)
          anova(model_TD_c, model_TD_d) # not significant so no interaction between
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
object	5	1184.056	1196.028	-587.0281	1174.056	NA	NA	NA
..1	6	1181.261	1195.628	-584.6304	1169.261	4.795271	1	0.02853797

```
In [148]: # Checking for interaction between CASE and FREQUENCY
          model_TD_e = lmer (TOTAL_DURATION~CASE + FREQUENCY + (1 | ITEM),
                               data=data, REML=FALSE)
          model_TD_f = lmer (TOTAL_DURATION~CASE * FREQUENCY + (1 | ITEM),
                               data=data, REML=FALSE)
          anova(model_TD_e, model_TD_f) # no significant interaction between CASE a
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
object	5	1194.352	1206.324	-592.1758	1184.352	NA	NA	NA
..1	6	1196.177	1210.544	-592.0887	1184.177	0.1741596	1	0.6764412

1.8.7 First-pass regression

```
In [149]: model_R_simple = lmer (FIRST_PASS_REGRESSION ~CASE + (1 | ITEM),
                                 data=data, REML=FALSE)
          # summary(model_R_simple)
```

```
In [150]: model_R_a = lmer (FIRST_PASS_REGRESSION ~CASE+FREQUENCY + (1 | ITEM),
                                 data=data, REML=FALSE)
```

```
In [151]: anova(model_R_simple, model_R_a)
          # not significant so FREQUENCY does not improve model fit
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
object	4	102.9840	112.5618	-47.49200	94.98401	NA	NA	NA
..1	5	104.9812	116.9534	-47.49058	94.98115	0.00285726	1	0.9573707

```
In [152]: model_R_b = lmer (FIRST_PASS_REGRESSION ~CASE+PREDICTABILITY + (1 | ITEM)
                                    data=data, REML=FALSE)
```

```
In [153]: anova(model_R_simple, model_R_b)
          # not significant so PREDICTABILITY does not improve model fit
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
object	4	102.9840	112.5618	-47.49200	94.98401	NA	NA	NA
..1	5	104.6987	116.6709	-47.34935	94.69870	0.2853102	1	0.5932416

```
In [154]: # Checking for interaction between CASE and PREDICTABILITY
          model_R_c = lmer (FIRST_PASS_REGRESSION~CASE + PREDICTABILITY + (1 | ITEM
                                  data=data, REML=FALSE)
          model_R_d = lmer (FIRST_PASS_REGRESSION~CASE * PREDICTABILITY + (1 | ITEM
                                  data=data, REML=FALSE)
          anova(model_R_c, model_R_d) #  not significant so there is no interaction
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
object	5	104.6987	116.6709	-47.34935	94.69870	NA	NA	NA
..1	6	105.9649	120.3316	-46.98246	93.96492	0.7337814	1	0.3916602

```
In [155]: # Checking for interaction between CASE and FREQUENCY
          model_R_e = lmer (FIRST_PASS_REGRESSION~CASE + FREQUENCY + (1 | ITEM),
                                  data=data, REML=FALSE)
          model_R_f = lmer (FIRST_PASS_REGRESSION~CASE * FREQUENCY + (1 | ITEM),
                                  data=data, REML=FALSE)
          anova(model_R_e, model_R_f) # no significant interaction between CASE and
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
object	5	104.9812	116.9534	-47.49058	94.98115	NA	NA	NA
..1	6	106.4134	120.7801	-47.20670	94.41340	0.5677537	1	0.4511529

1.8.8 Summary:

Eye movements by both participants are influenced by word frequency and contextual predictability, but in inconsistent ways. The neurologically healthy participant demonstrates a word frequency effect in the predicted direction for gaze duration (increase in gaze duration for low frequency words), and a predictability effect in the expected direction for total duration (prolonged total fixation durations on unpredictable words). The participant with aphasia showed a word frequency effect for gaze duration that was in the non-predicted direction (longer gaze duration for high frequency words), but a predictability effect for total fixation duration in the expected direction and in parallel to the neurologically healthy participant. Both participants seemed to be differently affected by word frequency and predictability with respect to first-pass regressions. The neurologically healthy particiapnts was more likely to regress out of high frequency words if they were unpredictable than low frequency words. The participant with aphasia, however, regressed more out of unpredictable low frequency words than unpredictable high frequency words. However, the models did not find that this difference between participants was significant.

```
In [ ]:
```