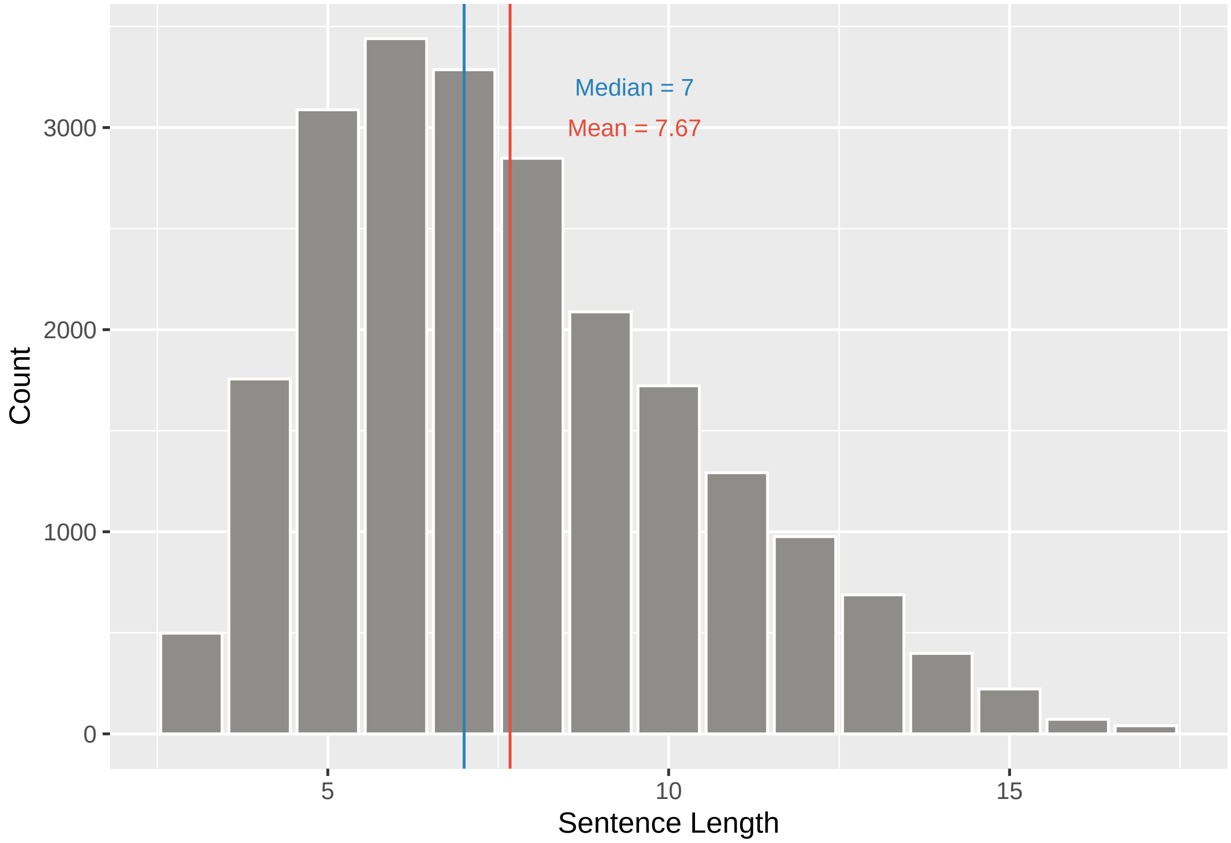
DLM during dialogue and DLM in speech vs written text

2023-08-11

RLAs Baselines of Dialogue Corpus Including Phase3 6 files(Token and disflueiencies removed)

1. Load the data



```
##      lang              dtype              sent_id          length.V1
##  Length:22414       Length:22414       Min.   :   1.0   Min.   :-1.699657
##  Class :character   Class :character   1st Qu.: 244.0   1st Qu.:-0.608662
##  Mode  :character   Mode  :character   Median : 481.0   Median :-0.244996
##                                        Mean   : 507.8   Mean   : 0.000000
##                                        3rd Qu.: 756.0   3rd Qu.: 0.482334
##                                        Max.   :1279.0   Max.   : 3.391657
##    avg_arity        max_arity         projD          avgHD
##  Min.   :0.6667   Min.   : 1.000   Min.   :2.000   Min.   :1.000
##  1st Qu.:0.8571   1st Qu.: 3.000   1st Qu.:3.000   1st Qu.:1.000
##  Median :0.8750   Median : 4.000   Median :3.000   Median :1.200
##  Mean   :0.8727   Mean   : 3.866   Mean   :3.386   Mean   :1.245
##  3rd Qu.:0.9000   3rd Qu.: 5.000   3rd Qu.:4.000   3rd Qu.:1.375
##  Max.   :0.9444   Max.   :11.000   Max.   :7.000   Max.   :3.083
##      avgDD
##  Min.   :1.000
##  1st Qu.:1.667
##  Median :2.000
##  Mean   :2.164
##  3rd Qu.:2.500
##  Max.   :7.562
```

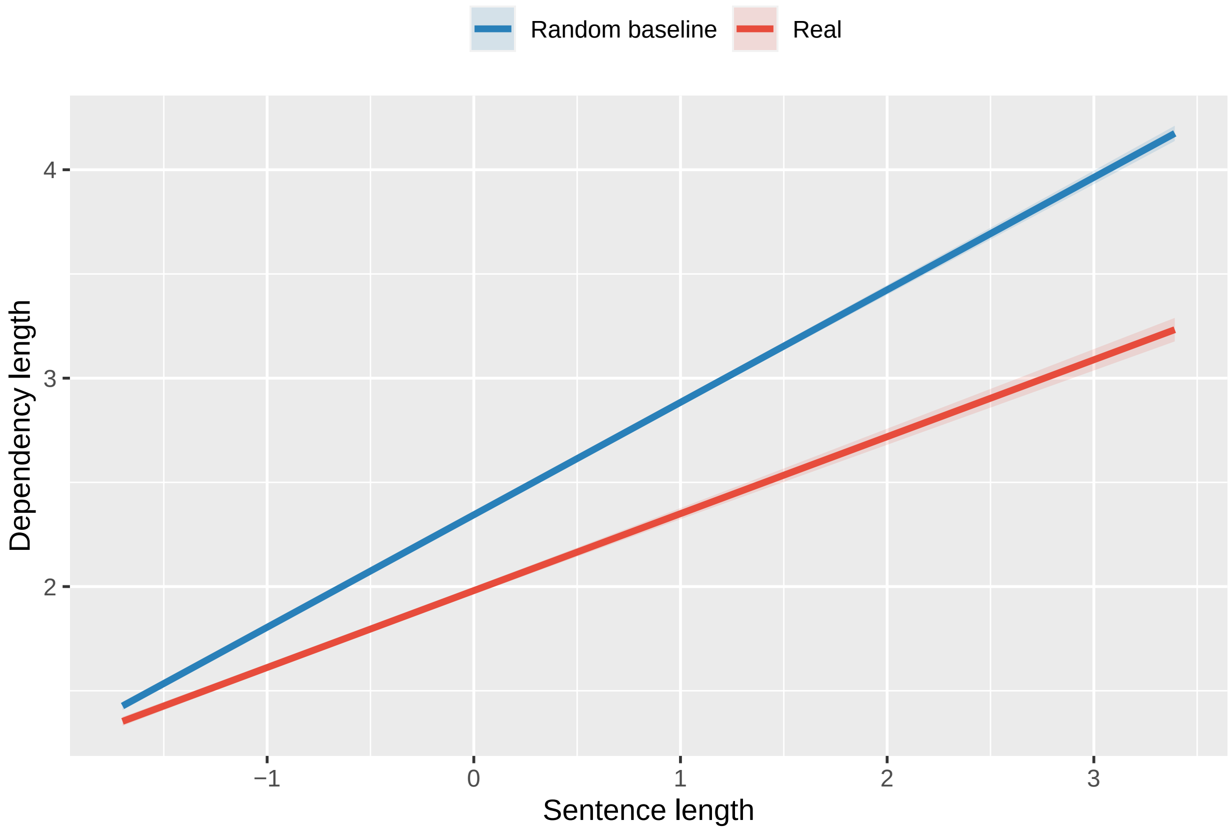
2. Fit the lmer model

```
#Running the model
#Avg Dependency Length ~ Sentence Length * Tree Type(Sentence Length * Tree Type | File ID)

m1.RLA <- lmer(avgDD~length*dtype+(length*dtype|lang),data=Data.RLA,control=lmerControl(optimizer="boby
summary(m1.RLA)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: avgDD ~ length * dtype + (length * dtype | lang)
##    Data: Data.RLA
## Control: lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e+05))
##
## REML criterion at convergence: 32570.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5741 -0.7016 -0.1296  0.5416  6.7297
##
## Random effects:
##  Groups   Name              Variance  Std.Dev. Corr
##  lang     (Intercept)       0.0003576 0.01891
##           length            0.0001103 0.01050   0.17
##           dtypereal         0.0024731 0.04973  -0.59  0.12
##           length:dtypereal  0.0005705 0.02389  -0.84  0.10  0.93
##  Residual                   0.2493218 0.49932
## Number of obs: 22414, groups:  lang, 30
##
## Fixed effects:
##                   Estimate Std. Error        df t value Pr(>|t|)
## (Intercept)       2.344248   0.005893 27.396794  397.79   <2e-16 ***
## length            0.539646   0.005141 37.666520  104.97   <2e-16 ***
## dtypereal        -0.363586   0.011365 27.380481  -31.99   <2e-16 ***
## length:dtypereal -0.170501   0.008088 35.502106  -21.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) length dtyprl
## length       0.047
## dtypereal   -0.618  0.031
## lngth:dtypr -0.277 -0.525  0.418
## optimizer (bobyqa) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

3. Plot fitted regression



RLAs New Distributed (Dialogue_Phase Vs Hindi_Text)

a. Load the Data

```
##      lang              dtype              sent_id          length
##  Length:2607        Length:2607        Min.   :    2   Min.   : 4.000
##  Class :character   Class :character   1st Qu.:  409   1st Qu.: 7.000
##  Mode  :character   Mode  :character   Median :  861   Median : 8.000
##                                        Mean   : 3076   Mean   : 8.133
##                                        3rd Qu.: 5304   3rd Qu.:10.000
##                                        Max.   :13295   Max.   :11.000
##    avg_arity        max_arity         projD           maxHD
##  Min.   :0.8000   Min.   :1.000   Min.   :3.000   Min.   :1.00
##  1st Qu.:0.8750   1st Qu.:2.000   1st Qu.:4.000   1st Qu.:2.00
##  Median :0.8889   Median :3.000   Median :4.000   Median :3.00
##  Mean   :0.8850   Mean   :3.044   Mean   :4.226   Mean   :2.74
##  3rd Qu.:0.9091   3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:3.00
##  Max.   :0.9167   Max.   :6.000   Max.   :9.000   Max.   :7.00
##      avgDD          Genre
##  Min.   :1.000   Length:2607
##  1st Qu.:1.750   Class :character
##  Median :2.100   Mode  :character
##  Mean   :2.154
##  3rd Qu.:2.472
##  Max.   :4.556
```

b. Fit the lm Model

```
#setting up sum contrast
contrasts(Data.m.RLA$dtype)
```

```
##              real_Dialouge real_Text
## random                   0         0
## real_Dialouge            1         0
## real_Text                0         1
```

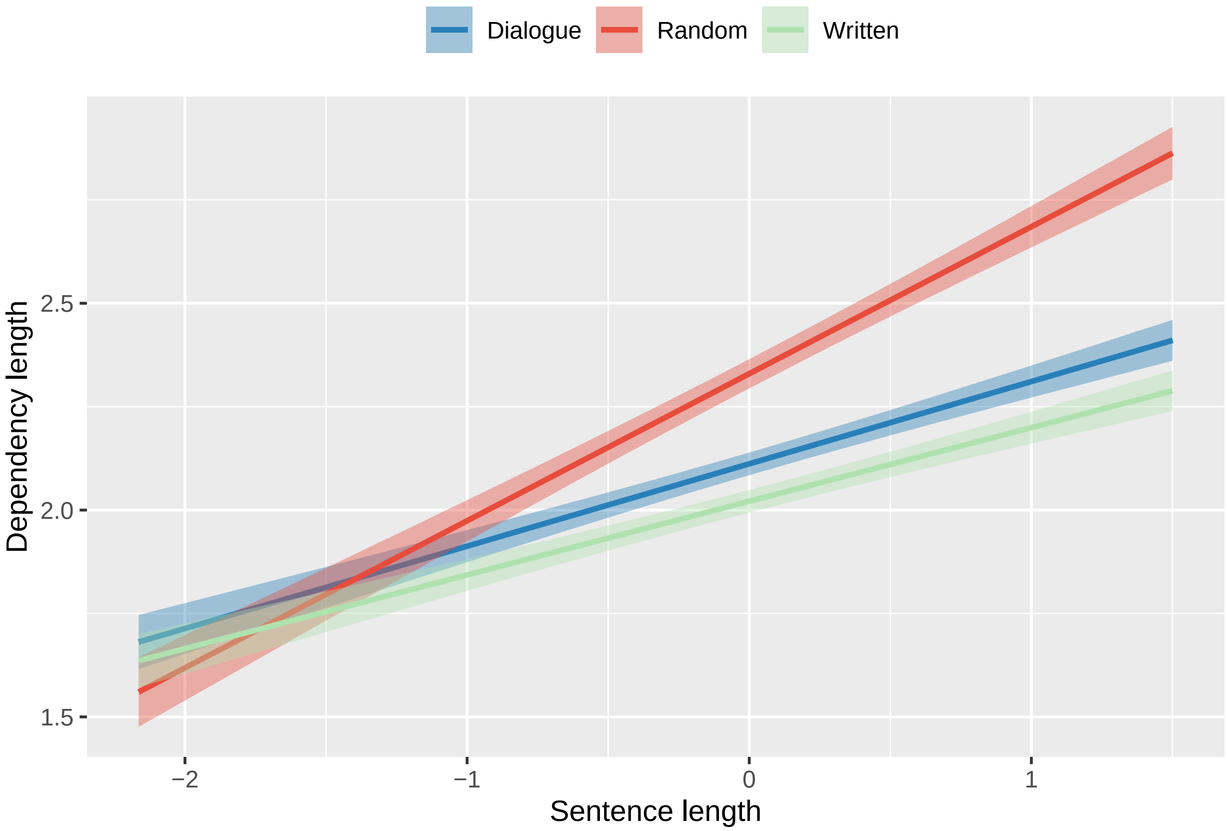
```
#Avg Dependency Length ~ Sentence Length * Tree Type
m1.m.RLA<- lm(avgDD~length*dtype, data = Data.m.RLA)

summary(m1.m.RLA)
```

```
##
## Call:
## lm(formula = avgDD ~ length * dtype, data = Data.m.RLA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.36306 -0.30637 -0.01897  0.27095  1.88956
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.32953    0.01540 151.238  < 2e-16 ***
## length                      0.35547    0.01541  23.073  < 2e-16 ***
## dtypereal_Dialouge         -0.21781    0.02178  -9.999  < 2e-16 ***
## dtypereal_Text             -0.30831    0.02178 -14.153  < 2e-16 ***
## length:dtypereal_Dialouge -0.15644    0.02179  -7.180 9.03e-13 ***
## length:dtypereal_Text     -0.17739    0.02179  -8.142 5.96e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4541 on 2601 degrees of freedom
## Multiple R-squared:  0.2865, Adjusted R-squared:  0.2852
## F-statistic: 208.9 on 5 and 2601 DF,  p-value: < 2.2e-16
```

c. Plot a graph

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```


Dialogue
Random
Written
Dependency length
2.5
2.0
1.5
−2
−1
0
1
Sentence length