Speech Technologies: Evaluation of SRILM (Graphics with R)

Initialisation

This file was generated using R Studio. In this document I explain in detail the steps done to obtain the plots shown in the Exercise document (paper format).

The first step, is to load the excel file and store it in dataframe format. Since we have two sheets, one for the training corpus descriptions and another with the test corpus results we will use two objects

```
library(gdata)
trainResults = read.xls(pathToFile, sheet=2)
summary(trainResults)
```

```
##          X         words             sentences
##  Europ v7 :1   Min.   :  8542120   Min.   : 391233
##  News 2007:1   1st Qu.: 42668980   1st Qu.:1761458
##  News 2010:1   Median : 65873400   Median :3000375
##  News Comm:1   Mean   : 69249730   Mean   :3297301
##                3rd Qu.: 92454150   3rd Qu.:4536218
##                Max.   :136710000   Max.   :6797220
```

```
testResults = read.xls(pathToFile, sheet=1)
summary(testResults)
```

```
##                X      X.1      sentences        words
##                 :16   CS:4   Min.   :1370   Min.   :23668
##  Europarl v7/v8 : 1   DE:4   1st Qu.:1500   1st Qu.:24204
##  New Crawl 2007 : 1   FI:4   Median :2169   Median :40771
##  New Crawl 2010 : 1   FR:4   Mean   :2103   Mean   :38547
##  News Commentary: 1   RU:4   3rd Qu.:2656   3rd Qu.:47160
##                              Max.   :2818   Max.   :56934
##       OOVs           logprob             ppl              ppl1
##  Min.   : 490.0   Min.   :-164388   Min.   : 425.9   Min.   : 586.3
##  1st Qu.: 832.5   1st Qu.:-137124   1st Qu.: 481.6   1st Qu.: 676.7
##  Median :1203.5   Median :-114856   Median : 600.6   Median : 915.2
##  Mean   :1536.0   Mean   :-109116   Mean   : 655.0   Mean   : 963.7
##  3rd Qu.:2470.8   3rd Qu.: -67975   3rd Qu.: 795.0   3rd Qu.:1151.3
##  Max.   :3250.0   Max.   : -66743   Max.   :1136.2   Max.   :1741.2
```

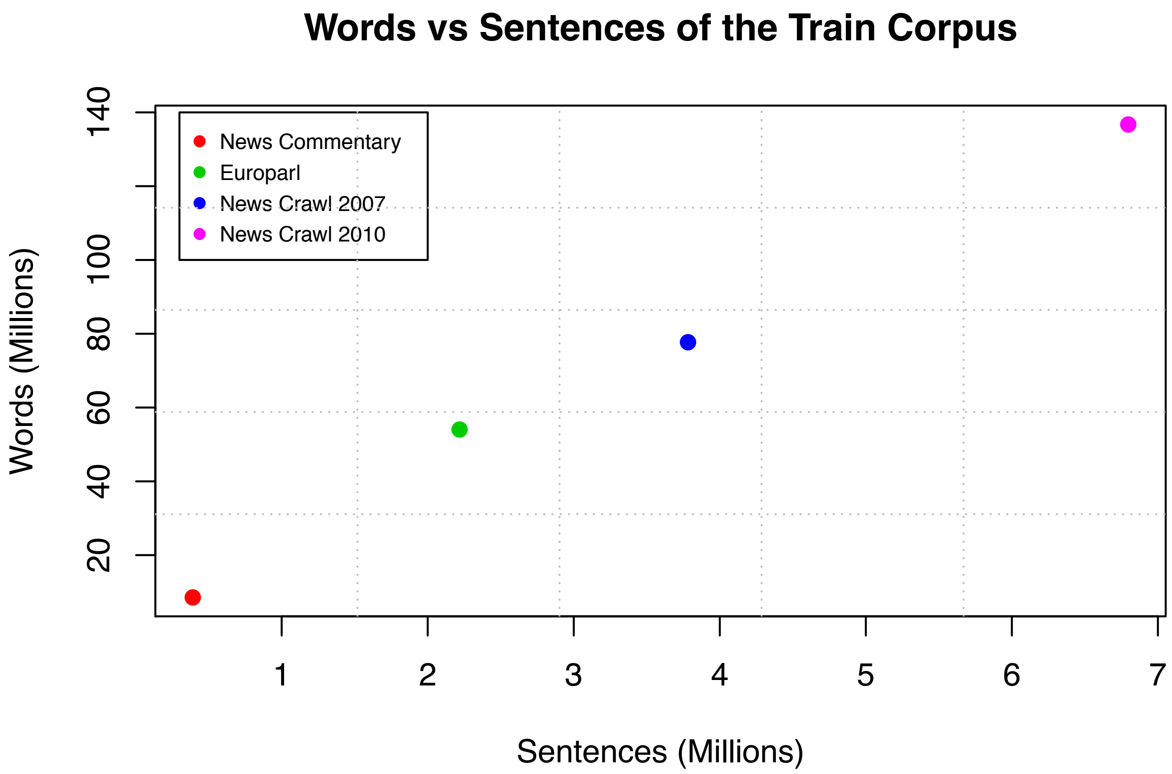
Before going into the plots, we initialise the formatting, that is the colors and shapes that will be used in subsequent plots to differentiate the samples.

```
train_col=c(2,3,4,6) #,c(5,5,5)
train_pch=rep(19,4)
test_col=rep(1,5)
test_pch=c(8,15,17,18,19) #,3
```

Train Corpus

Let us now plot the relation of the words and sentences of the training corpus

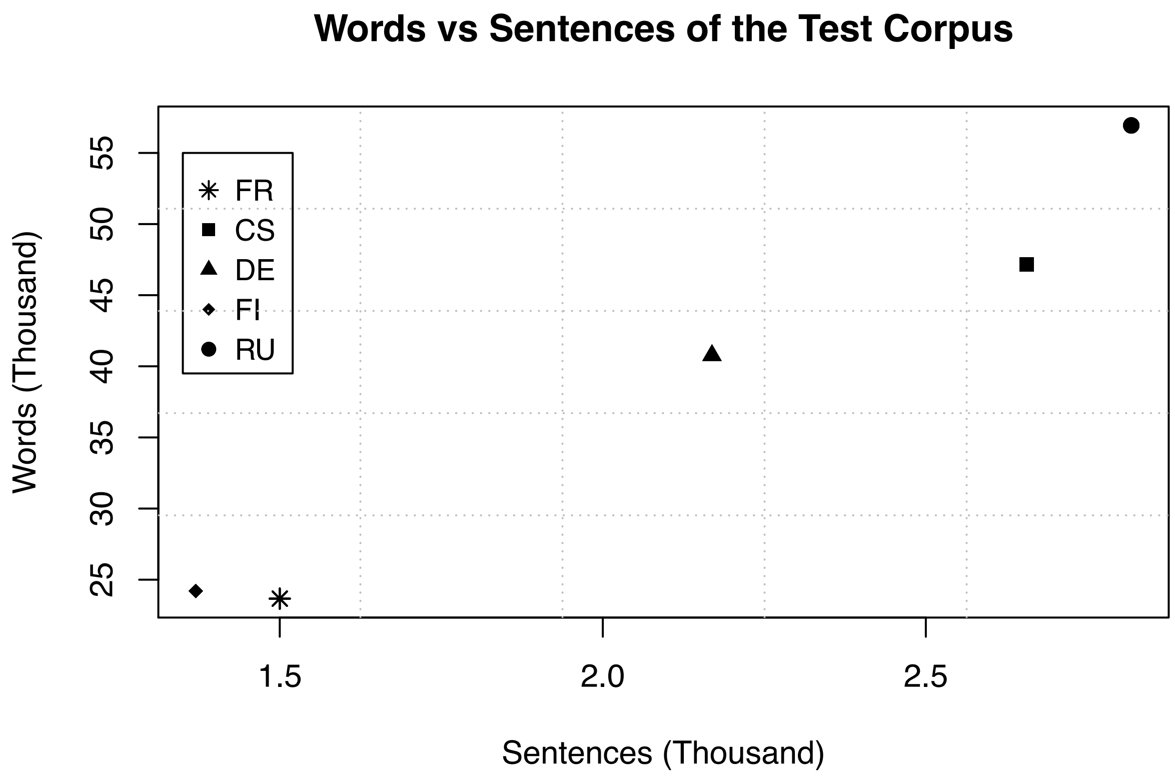
```
plot(words~sentences, data=trainResults/1e6, pch=train_pch, col=train_col,
     ylab="Words (Millions)", xlab="Sentences (Millions)",
     main="Words vs Sentences of the Train Corpus")
legend(x = c(0.3,2), y= c(140,100), inset=c(0.1,0), pch = train_pch, col=train_col,  c("News Commentary
       cex = 0.7)
grid(5, 5, col = "gray", lty = "dotted",
     lwd = par("lwd"), equilogs = TRUE)
```



Test Corpus

Let us now plot the relation of the words and sentences of the test corpus

```
plot(words~sentences, data=testResults/1e3, pch=test_pch, col=test_col,
     ylab="Words (Thousand)", xlab="Sentences (Thousand)",
     main="Words vs Sentences of the Test Corpus")
legend(x = c(1.350,1.520), y= c(55.000,39.500), inset=c(0.1,0), pch = test_pch, col=test_col,  c("FR",
       cex = 0.9)
grid(5, 5, col = "gray", lty = "dotted",
     lwd = par("lwd"), equilogs = TRUE)
```



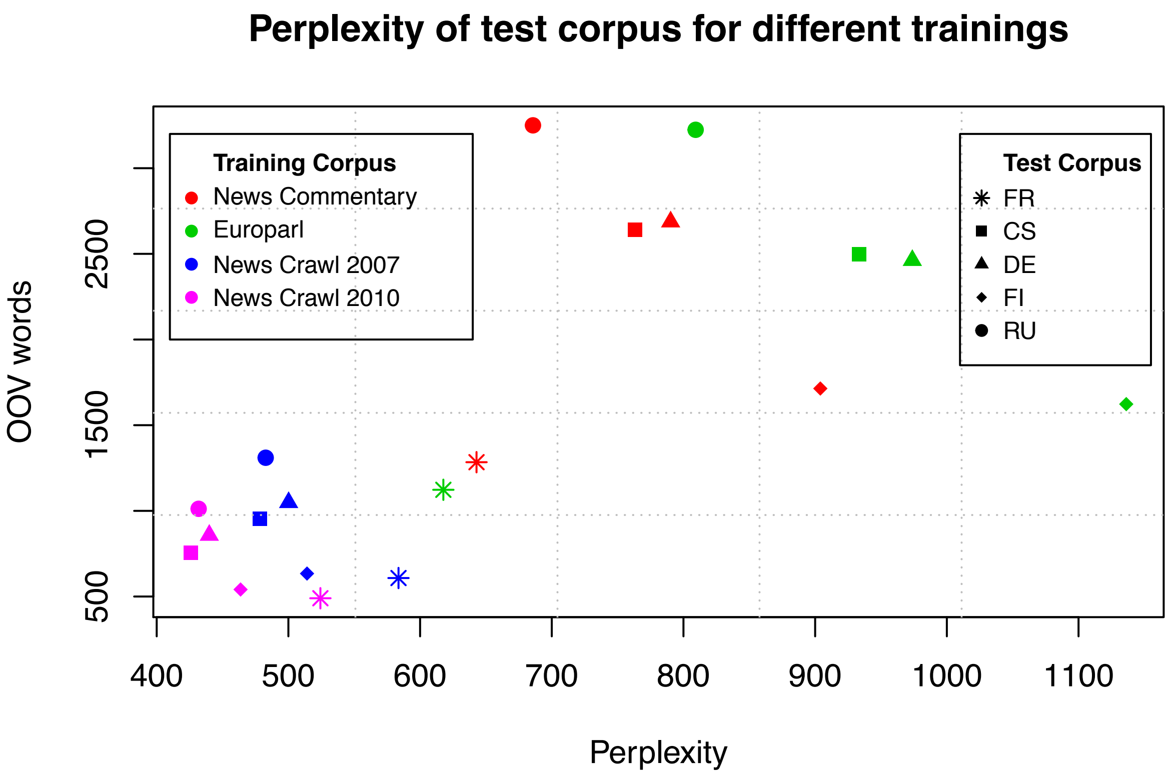
Final Results

And finally, this are the final results

```
# Plot results for OOVs~ppl
plot(OOVs~ppl, data=testResults, pch=rep(test_pch,4), col=rep(train_col,c(5,5,5,5)),
            ylab="OOV words", xlab="Perplexity",
            main="Perplexity of test corpus for different trainings")

# Add grid
grid(5, 5, col = "gray", lty = "dotted",
     lwd = par("lwd"), equilogs = TRUE)

# Add legend of the colors (training sets)
legend(x = c(410,640), y= c(3200,2000), inset=c(0.2,0), col=c(0,train_col), pch = 19,  c(expression(bol
      cex = 0.75)
# Add legend of the shapes (test sets)
legend(x = c(1010,1155), y= c(3200,1850), inset=c(0.1,0), col= c(0,test_col), pch = c(0,test_pch),  c(e
       cex = 0.75)
```



Perplexity of test corpus for different trainings
Training Corpus
News Commentary
Europarl
News Crawl 2007
News Crawl 2010
Test Corpus
FR
CS
DE
FI
RU
OOV words
500
1500
2500
400
500
600
700
800
900
1000
1100
Perplexity