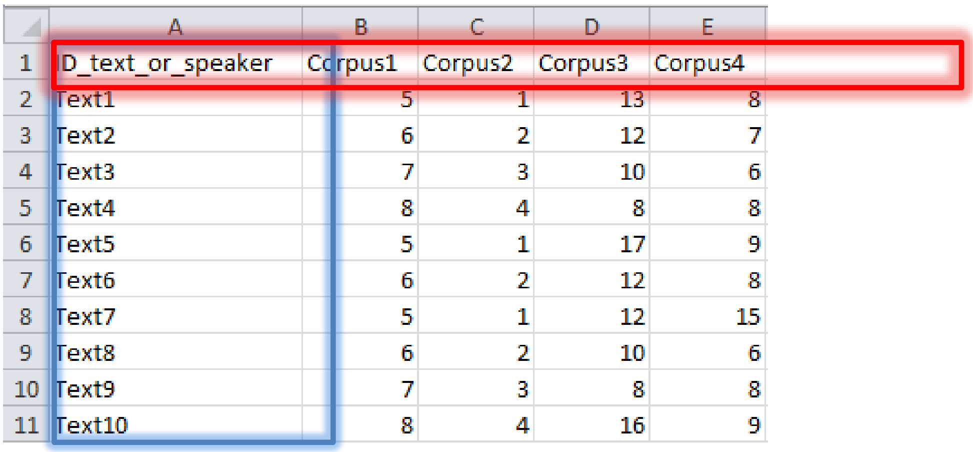
Graph tool instructions and R code

1) Prepare data: tab-delimited format

Data need to be inputted in a tab-delimited format. This can be easily achieved by preparing the data in a spread sheet program such as Excel or Calc. There must be no spaces or hyphens in variable names or values; this means that e.g. Text1 or Text_1 are acceptable values but Text 1 or Text-1 are not.



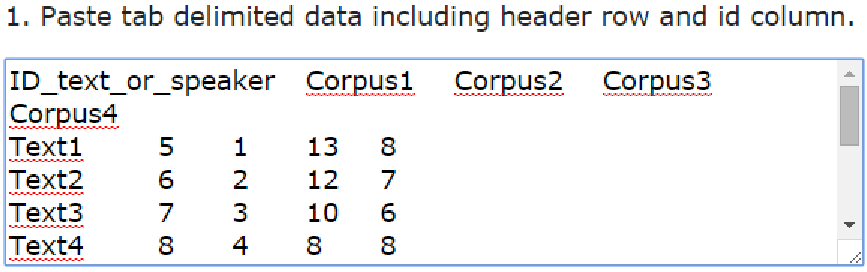
	A	B	C	D	E
1	ID_text_or_speaker	Corpus1	Corpus2	Corpus3	Corpus4
2	Text1	5	1	13	8
3	Text2	6	2	12	7
4	Text3	7	3	10	6
5	Text4	8	4	8	8
6	Text5	5	1	17	9
7	Text6	6	2	12	8
8	Text7	5	1	12	15
9	Text8	6	2	10	6
10	Text9	7	3	8	8
11	Text10	8	4	16	9

header row with the names of (sub)corpora

text or speaker IDs

2) Input data: copy-paste

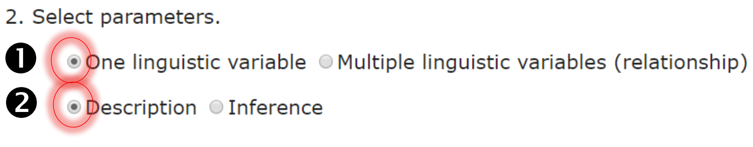
Simply copy-paste the data including the header row and ID column from Excel or Calc in the text box.



1. Paste tab delimited data including header row and id column. For help click here.

```
ID_text_or_speaker   Corpus1   Corpus2   Corpus3
Corpus4
Text1      5     1     13    8
Text2      6     2     12    7
Text3      7     3     10    6
Text4      8     4     8     8
```

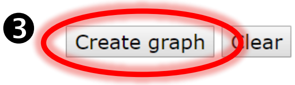
3) Select parameters and create graph



2. Select parameters.

❶ ● One linguistic variable ○ Multiple linguistic variables (relationship)

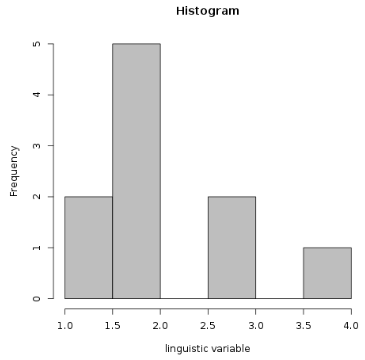
❷ ● Description ○ Inference



Types of graphs produced

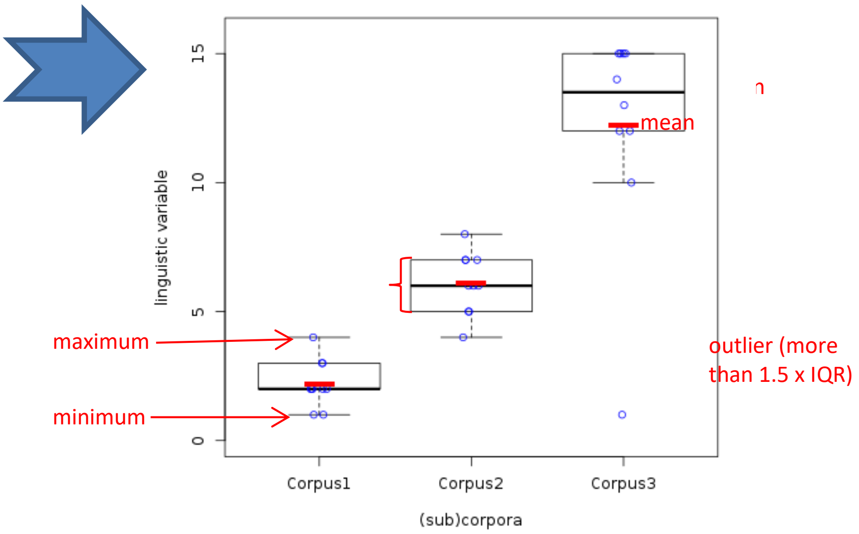
A) Histogram – One linguistic variable, Description, 1 corpus, multiple texts/speakers

Text	Corpus1
text1	1
text2	2
text3	1
text4	3
text5	3
text6	2
text7	1
text8	1
text9	2
text10	2



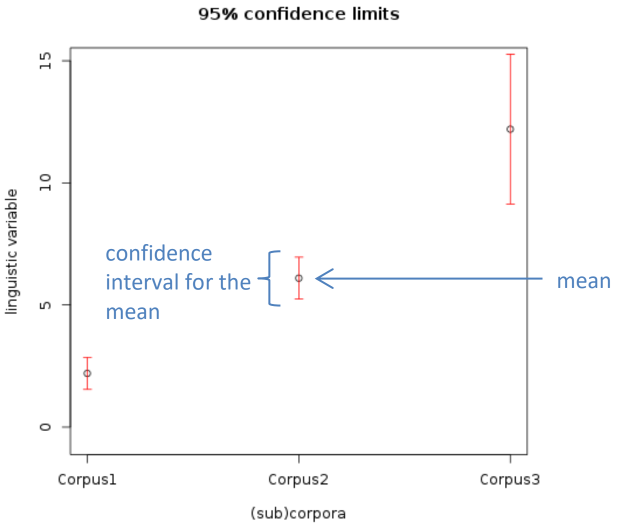
B) Boxplot – One linguistic variable, Description, multiple corpora, multiple texts/speakers

Text	Corpus1	Corpus2	Corpus3
text1	2	5	10
text2	2	5	15
text3	1	6	15
text4	3	7	14
text5	3	6	13
text6	2	7	12
text7	4	4	15
text8	1	7	15
text9	2	6	12
text10	2	8	1

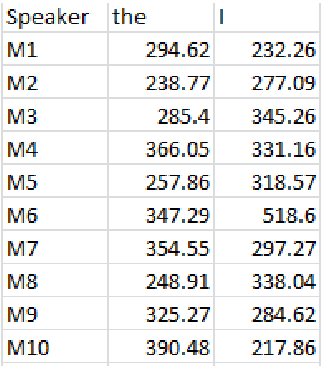


C) Error bars: 95% Confidence interval(s) – One linguistic variable, Inference, 1 corpus/multiple corpora, multiple texts/speakers

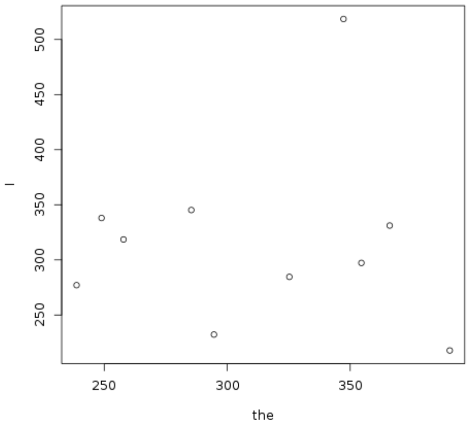
Text	Corpus1	Corpus2	Corpus3
text1	2	5	10
text2	2	5	15
text3	1	6	15
text4	3	7	14
text5	3	6	13
text6	2	7	12
text7	4	4	15
text8	1	7	15
text9	2	6	12
text10	2	8	1



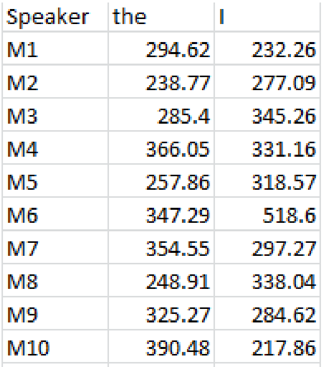
D) Scatter plot – Multiple linguistic variables, Description, 1 corpus, multiple texts/speakers



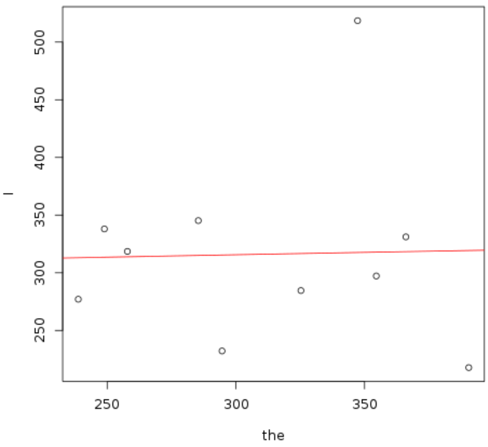
Speaker	the	I
M1	294.62	232.26
M2	238.77	277.09
M3	285.4	345.26
M4	366.05	331.16
M5	257.86	318.57
M6	347.29	518.6
M7	354.55	297.27
M8	248.91	338.04
M9	325.27	284.62
M10	390.48	217.86



E) Scatter plot with regression line – Multiple linguistic variables, Inference, 1 corpus, multiple texts/speakers

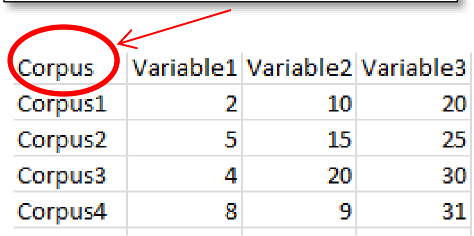


Speaker	the	I
M1	294.62	232.26
M2	238.77	277.09
M3	285.4	345.26
M4	366.05	331.16
M5	257.86	318.57
M6	347.29	518.6
M7	354.55	297.27
M8	248.91	338.04
M9	325.27	284.62
M10	390.48	217.86

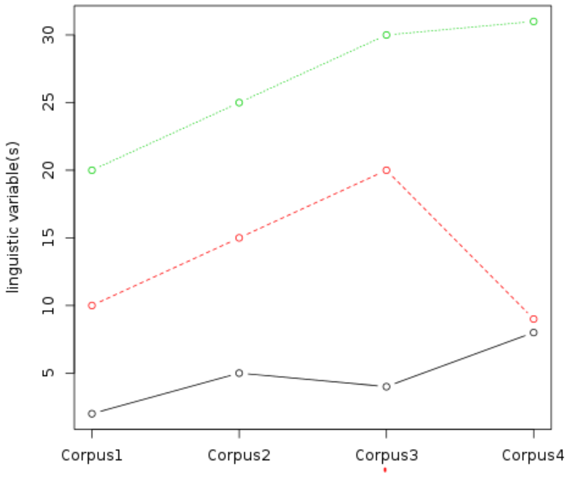


F) Line chart – One linguistic variable, Description, 1 corpus/multiple corpora, single value per corpus

The word ‘corpus’ or ‘corpora’ needs to be specified in column 1, row 1.

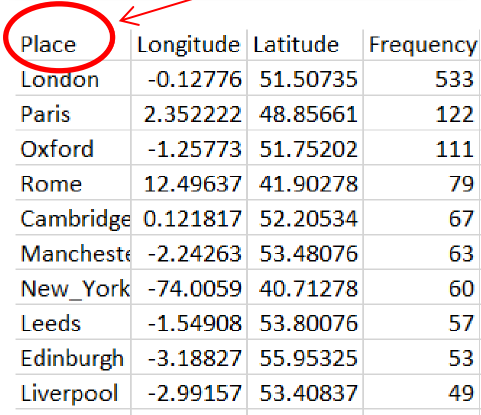


Corpus	Variable1	Variable2	Variable3
Corpus1	2	10	20
Corpus2	5	15	25
Corpus3	4	20	30
Corpus4	8	9	31



G) Geomapping – One linguistic variable, Place, Longitude, Latitude and Frequency

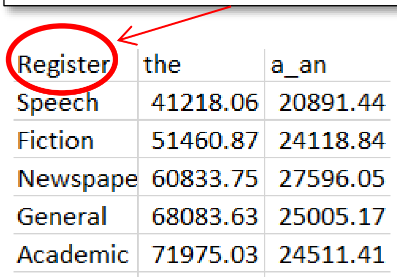
The word ‘place(s)’ or ‘location(s)’ needs to be specified in column 1, row 1.



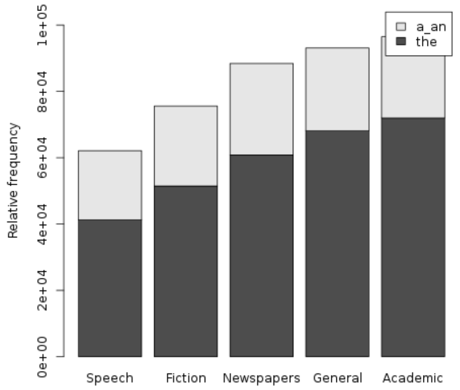
Place	Longitude	Latitude	Frequency
London	-0.12776	51.50735	533
Paris	2.352222	48.85661	122
Oxford	-1.25773	51.75202	111
Rome	12.49637	41.90278	79
Cambridge	0.121817	52.20534	67
Mancheste	-2.24263	53.48076	63
New_York	-74.0059	40.71278	60
Leeds	-1.54908	53.80076	57
Edinburgh	-3.18827	55.95325	53
Liverpool	-2.99157	53.40837	49

H) Stacked barchart – One linguistic variable, Place, Longitude, Latitude and Frequency

The word ‘genre’ or ‘register’ or ‘text_type’ needs to be specified in column 1, row 1.

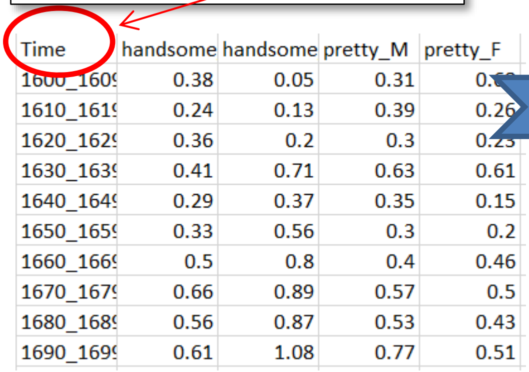


Register	the	a_an
Speech	41218.06	20891.44
Fiction	51460.87	24118.84
Newspape	60833.75	27596.05
General	68083.63	25005.17
Academic	71975.03	24511.41

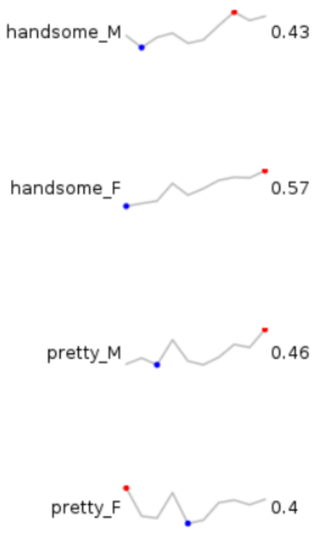


I) Sparklines – One/many linguistic variable(s), series, relative frequencies of one or up to four linguistic variables

The word ‘spark’ or ‘sparkline(s)’ or ‘time’ or ‘series’ needs to be specified in column 1, row 1.

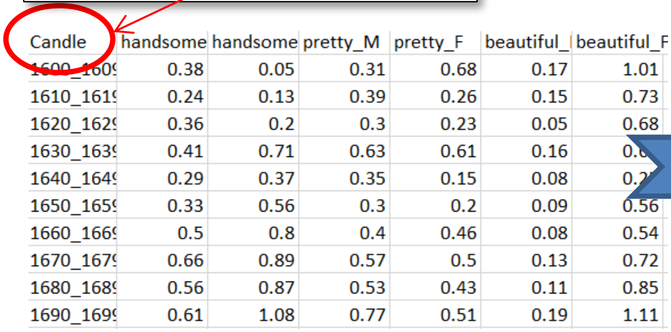


Time	handsome	handsome	pretty_M	pretty_F
1600_1609	0.38	0.05	0.31	0.69
1610_1619	0.24	0.13	0.39	0.26
1620_1629	0.36	0.2	0.3	0.23
1630_1639	0.41	0.71	0.63	0.61
1640_1649	0.29	0.37	0.35	0.15
1650_1659	0.33	0.56	0.3	0.2
1660_1669	0.5	0.8	0.4	0.46
1670_1679	0.66	0.89	0.57	0.5
1680_1689	0.56	0.87	0.53	0.43
1690_1699	0.61	1.08	0.77	0.51

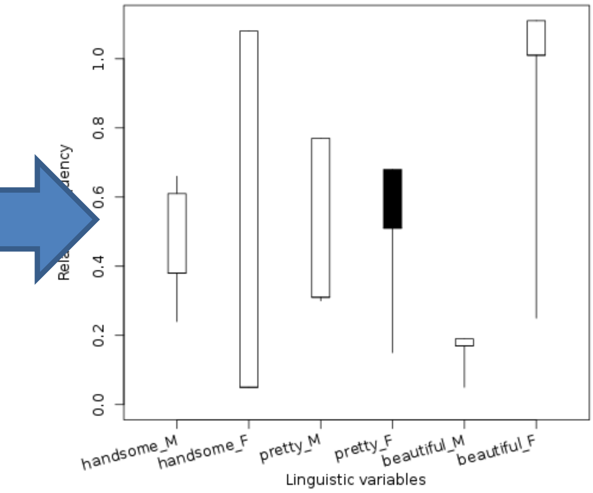


J) Candlestick plot – One/many linguistic variable(s), series, relative frequencies of one or up to four linguistic variables

The word ‘candle or ‘candlestick’ needs to be specified in column 1, row 1.



Candle	handsome	handsome	pretty_M	pretty_F	beautiful_	beautiful_F
1600_1609	0.38	0.05	0.31	0.68	0.17	1.01
1610_1619	0.24	0.13	0.39	0.26	0.15	0.73
1620_1629	0.36	0.2	0.3	0.23	0.05	0.68
1630_1639	0.41	0.71	0.63	0.61	0.16	0.6
1640_1649	0.29	0.37	0.35	0.15	0.08	0.2
1650_1659	0.33	0.56	0.3	0.2	0.09	0.56
1660_1669	0.5	0.8	0.4	0.46	0.08	0.54
1670_1679	0.66	0.89	0.57	0.5	0.13	0.72
1680_1689	0.56	0.87	0.53	0.43	0.11	0.85
1690_1699	0.61	1.08	0.77	0.51	0.19	1.11



code

```
Datasets are available at:
http://corpora.lancs.ac.uk/stats/data/graph_tool_examples.csv

http://corpora.lancs.ac.uk/stats/data/graph_tool_examples.xlsx

#histogram
hist(x, breaks="Sturges", col="gray", xlab="linguistic variable",
main="Histogram")


#boxplot with points and mean overlay
boxplot(myData, ylab = "linguistic variable",xlab="(sub)corpora", outline =
FALSE, ylim=c(0, max(myData, na.rm=TRUE)*1.05)); i = 1;while(i <=
ncol(myData)) { for(v in myData[,i]){points(jitter(i,3/i),v, col = "blue",
pch=1, cex = 1)};
points(i, mean(myData[,i],trim = 0, na.rm = TRUE), col = "red", pch="_",
cex = 4) i= i+1; }


#scatter plot with regression line
plot(myData); fitline <- lm(myData[,2] ~ myData[,1]);
abline(fitline,col="red")


#error bars
error.bars(myData,stats=NULL, ylab = "linguistic
variable",xlab="(sub)corpora", main=NULL,eyes=FALSE, ylim = NULL,
xlim=NULL,alpha=.05,sd=FALSE, labels = NULL, pos = NULL, arrow.len =
0.05,arrow.col="red", add = FALSE,bars=FALSE,within=FALSE, col="red")


#line chart
matplot(myData, type = c("myData"),pch=1, ylab="linguistic variable(s)",
xaxt = "n", col = 1:4); axis(1, at=x, labels=v)
#stacked barchart
matplot(myData, type = c("myData"),pch=1, ylab="linguistic variable(s)",
xaxt = "n", col = 1:4); axis(1, at=x, labels=v)


#geomapping
library(maps); library(mapdata)#load libraries
map('worldHires',xlim=c((min(myData[,1])-5),(max(myData[,1])+5)),
ylim=c((min(myData[,2])-5),(max(myData[,2])+5)));
i=1;while(i<=length(myData[,1])){points(myData[i,1],myData[i,2],
col=2,pch=19, cex=(4*(myData[i,3]/x)+0.5)); i<-i+1;};")

#sparklines
par(mfrow=c(4,1),mar=c(5,7,4,2),omi=c(0.2,2,0.2,2)); for(i in
1:4){x=round(mean(b[,i]),2);plot(b[,i],ann=FALSE,axes=FALSE,type="l",col="g
ray",lwd=2); mtext(side=2,at=x,names(b[i]),las=2,col="black");
mtext(side=4,at=x,x,las=2,col="black");
points(which.min(b[,i]),min(b[,i]),pch=19,col="blue");
points(which.max(b[,i]),max(b[,i]),pch=19,col="red");}

#candlestick plot
      #prepare data
      min<-apply(s, 2, min)
      min<-as.vector(min)
      max<-apply(s, 2, max)
      max<-as.vector(max)
      first<-head(s,1)
      first<-as.numeric(first[1,])
      last<-tail(s,1)
      last<-as.numeric(last[1,])
      int_1<-ifelse(first<=last,first,last)
      int_2<-ifelse(last>=first,last,first)
      dir<-ifelse(first<=last,1,0)
      dir<-as.vector(dir)
      item<-colnames(s)
      item<-as.vector(item)
      order<-seq_along(item)
      data <- data.frame(order, item, min, int_1, int_2, max, dir)

      #create plot
      with(data,symbols(order, (int_1+ int_2)/2, boxplots=cbind(.25,int_2-
      int_1, int_1-min,max-int_2,0),inches=F,ylim=range(data[,-
      c(1:2)]),xaxt="n",ylab="Relative frequency", xlab="Linguistic
      variables", bg = ifelse(dir==0, "black", "white"))); data<-
      data[with(data, order(order)),];
      axis(1,seq_along(data$item),labels=FALSE);
      text(x=seq_along(data$item), y=par()$usr[3]-0.05*(par()$usr[4]-
      par()$usr[3]), labels=data$item, srt=15, adj=1, xpd=TRUE);
```