Course on R for linguists, part 2

Organizers: Bjørn-Helge Mevik & Diana Santos

        Computing Research Services, USIT

Date: 15 November, 10-13,

     17 November, 9-12,

Place: Henrik Wergelands Hus, PC stue 301

Basic instructions for the practical hands-on part

1.   Input the data: first cursory examination
2.   Exploring the data
3.   Data manipulation and clean-up
4.   Using a contingency table analysis
5.   Logistic regression example

1. Input the data: first cursory examination

Fetch the following datsets to R

```
> simple <- read.table("f:/R_kurs_BHM/Other/simple.txt",
header=TRUE)
```

```
> uhm <- read.table("f:/R_kurs_BHM/Gries/03-1_uh(m).txt",
header=TRUE)
```

```
> reaction <- read.table("f:/R_kurs_BHM/Gries/03-2-
3_reactiontimes.txt", header=TRUE)
```

```
> survey <- read.table("f:/R_kurs_BHM/Other/ots.txt", header=TRUE)
```

```
> vocdata <-
read.delim("f:/R_kurs_BHM/KeithJohnson/5sociolinguistics/Robins_dat
a.txt")
```

```
> dative <- read.table("f:/R_kurs_BHM/Gries/05-
4_dativealternation2.txt",header=TRUE)
```

Use the commands **dim**, **summary, names** to learn about the datasets. For example:

```
> dim(survey)
```

```
> names(dative)
```

```
> summary(simple)
```

Check whether you have to redefine/improve on the data as far as classes/categories are concerned. Are there any factors that were misunderstood as numeric values?

Explore the columns with **table**.

```
> table(survey$kjonn)
```

```
> table(simple$variety)
```

If you have time,

create new dataframes with a subset of the values of the ones you have,
```
> onlyfb<-subset(simple,theme=="football",(4:5))
> onlyfb<-subset(simple,theme=="football",c(2,4))
```
and save them
```
> write.table(onlyfb, file="fbcol.txt")
```

trim your dataset in order to only include a subset of the columns
```
> sub_simple<-simple[c(3,5)]
```

2. Exploring the data

Use some basic statistic measures to take the pulse on the data:
mean, sd, var, median, min, ...

Plot the several data with **assocplot, spineplot, barplot, barplot(table...), hist, boxplot**

Look carefully at the dative data set. What is wrong?

Hint: plot every column...
```
> plot(dative$REC_ACT)
> plot(dative$PAT_ACT)
```

Fix it, or fetch
```
> dative <- read.table("f:/R_kurs_BHM/Gries/05-
4_dativealternation.txt", header=TRUE)
```

to proceed with a correct dataset...

Try now to identify whether the length of the uhm dataset is significantly different between males or females

```
attach(uhm)
boxplot(LENGTH)
plot(LENGTH ~ SEX)
```

Can you say something from visual inspection? Do a t-test as well:
```
t.test(LENGTH ~ SEX)
```

Now visualize other issues

```
barplot(table(FILLER))
```

```
barplot(table(FILLER, SEX))
assocplot(table(FILLER, SEX))
```

and save the figures in different formats easier to include in
word processing programs.

```
> png("ola.png")
> assocplot(table(FILLER, SEX))
> devlist()
> devoff()
```

3. Data manipulation and clean-up

Now start working with the survey dataframe.

```
> attach(survey)
```

Convert the gender values 1 and 2 to clearer ids. (Hint: use
another column, and use descriptions such as M and F, or K and M)
One solution

```
> survey$gender[survey$kjonn=="1"] <- "M"
> survey$gender[survey$kjonn=="2"] <- "F"
> survey$gender<-factor(survey$gender)
```

Change the several appropriate values of language (ditt_sprog) to
«norsk»

Plot number of languages as a function of gender. Is the following
right?

```
> plot(andre_sprog~gender)
```

What is wrong? What can you do? Try different ways of showing the
distribution.

```
> spineplot(andre_sprog~gender)
```

```
> detach(survey)
```

4. Using a contingency table analysis

Now look at the vocalization data:

```
> vocdata$newage<-factor(vocdata$age,levels=c(1,2,4,5),
labels=c("teens","twenties","forties","fifties"))
```

```
> vocdata$lvoc<-
factor(vocdata$l,levels=c(1,3),labels=c("unvocalized","vocalized"),
exclude=c(2))
```

What did the previous commands do?
Try to create other columns with other names.

3

```
> attach(vocdata)
> table(lvoc,newage)
```

Do contingency analysis:

```
> summary(table(lvoc,newage))
> detach(vocdata)
```


5. Logistic regression example

Let us now do a logistic regression model to see which factors influence the use of the two dative constructions in English.

```
> DatAlt<-dative[complete.cases(dative),]
> attach(DatAlt)
> model.glm<-
glm(CONSTRUCTION1~V_CHANGPOSS*AGENT_ACT+V_CHANGPOSS*REC_ACT,
family=binomial)

> install.packages("car")
> library(car)
> Anova(model.glm)

> detach(DatAlt)
```

References for further work

Crawley, Michael J. *Statistics: An Introduction using R*. John Wiley & Sons, 2005

Gries, Stefan Th. "Some proposals towards more rigorous corpus linguistics". *Zeitschrift für Anglistik und Amerikanistik* **54**.2, 2006, pp. 191-202. http://www.linguistics.ucsb.edu/faculty/stgries/research/MoreRigCorpLing_ZAA.pdf

Gries, Stefan Th. "Exploring variability within and between corpora: some methodological considerations". *Corpora* **1**.2, 2007, pp. 109-51. http://www.linguistics.ucsb.edu/faculty/stgries/research/ExploringVariability_Corpora.pdf

Gries, Stefan Th. *Statistics for Linguistics with R: A Practical Introduction*. [Trends in Linguistics. Studies and Monographs [TiLSM] 208] Mouton de Gruyter, 2009.

Johnson, Keith. *Quantitative Methods in Linguistics*, Wiley-Blackwell, 2008.

Stefanowitsch, Anatol & Stefan Th. Gries. "Collostructions: investigating the interaction between words and constructions". *International Journal of Corpus Linguistics* **8**(2), pp. 209-243.

Stoll, Sabine & Stefan Th. Gries. "An association strength approach to characterizing development in corpora". *Journal of Child Language* **36**, 2009, pp. 1075-1090. http://www.eva.mpg.de/lingua/staff/stoll/pdf/Stoll&Gries_JCL.pdf