A Practical Course in Corpus Linguistics for Students with a Humanist Background

Mihaela Vela & Hannah Kermes
Language Science and Technology
Saarland University

Overview

- Practical course on Corpus Linguistics
- BA Language Science
  - Students with humanist background
  - Translatology and languages studies
  - Little or no experience in NLP

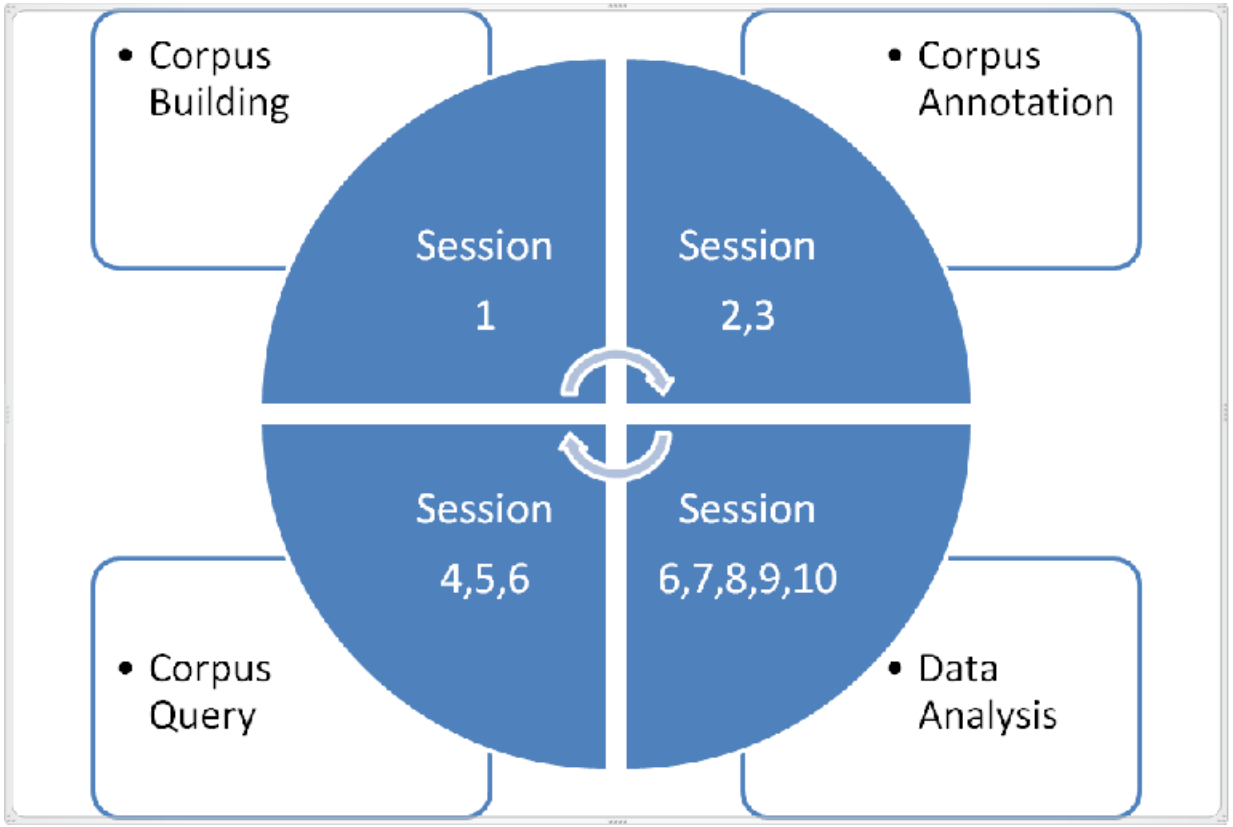
Challenges

- Students
  - Learning a totally new subject
  - Dealing with and solving technical problems
  - Coping with the demands of active learning
- Teachers
  - Motivating students by lowering the psychological and practical barriers
  - Avoiding or solving technical problems
  - Dealing with heterogeneous groups
  - Keeping track of learning success
  - Adapting to specific needs

General Concept

- Necessary skills and knowledge for empirical studies
- Constructed like a sample study
  - Tutorials representing single steps
- Applicable to different settings and target groups
- Active and collaborative learning
- Teacher as a moderator and assistant

Structure of the Course



Method, Tools and Data

- Method
  - Tutorial vs. exercise
  - Active learning in class vs. self learning
  - R Markdown
  - Course material on-line
- Tools
  - TreeTagger (Schmid, 1994)
  - CQPWeb (Hardie, 2012)
  - WebLicht (Hinrichs et al., 2010)
  - Excel/Libre Office
  - Notepad++
  - RStudio
- Data
  - RSC (Kermes et al., 2016)
  - Brown family (Brown (Francis and Kučera, 1979), Frown (Mair, 1999), etc)

Corpus Building

- Session 1
  - Corpus building with XML and TEI

A first exercise

Exercise

- create a `data` directory in the course directory on your USB-Stick
- download the file RSC_1009222.txt and RSC_1009222.pdf and
- save it in the `data` directory
- open the save file `RSC_1009222.txt` in `notepad++`
- the file is a full text version (plain text) of RSC_1009222.pdf
- full text has its advantages, but some information of the pdf file is lost
- mark the following things in the plain text file
  - title, paragraphs and sentences

Corpus Annotation

- Session 2
  - Tagging with the TreeTagger
  - Part-of-speech tagging of .txt and .xml files

Tagging text with XML/SGML tags

- the sample file `grimm_sample.txt` contained plain-text only
- the sample file `grimm_sample.xml` additionally contains meta-data information and annotations using `XML/SGML` tags
- What happens if you tag the text `grimm_sample.xml` with the same settings we used for `grimm_sample.txt` ? - give it a try!
  - the `XML/SGML` tags are treaded by TreeTagger as if they were normal words and are assigned a part-of-speach tag and a lemma
- however, what we want TreeTagger to do is ignore the `XML/SGML` tags, leaving them as they are
- in order to tell TreeTagger to ignore the `XML/SGML` tags, we need to tick the `Input Option` `SGML tags present`
  - the `XML/SGML` tags have to be on a separate line!
- tag `grimm_sample.xml` with this option and have a look at the output file.

Corpus Annotation

- Session 3
  - Corpus annotation with WebLicht
    - Additonal annotation layers
    - Processing chain with at least a tokenizer and the TreeTagger
  - Tokenization
  - Lemmatization
  - Pos-tagging
  - Parsing

Corpus Query

- Session 4
  - Regular expressions in Notepad++
  - Introduction to CQPWeb
- Session 5
  - Formulating patterns in CQPWeb

```
[word="search.+"]
[word="search.{2}"]
[word="search.{2,3}"]
```

preposition followed by `any` or `every` followed by a noun in singular

```
[pos= "IN"] "any|every" [pos= "NN"]
```

token with lemma `go` followed by `and` and any another word

```
[lemma= "go"] "and" []
```

Corpus Query & Data Analysis

- Session 6:
  - Data extraction and data formats
  - Manipulating CQPWeb query results

Extract the data set

To recall the parameters of our example study:

- **research question**: distribution of full verbs and their parts-of-speech across registers in the Brown corpus
- **query**: `[pos="VV.*"]`.
- **observation**: each instance of a full verb in the Brown corpus
- **features/variables**: verb lemma, part-of-speech of verb, register

Parameters for the custom tabulation:

- three columns (one for each variable, order does not matter)
  - column 1: verb lemma; attribute: `lemma` ; anchor: `match`
  - column 2: pos of verb; attribute: `pos` ; anchor: `match`
  - column 3: register; attribute: `text_reg` ; anchor: `match`
- output format: `simple tabulate output`

Data Analysis

- Session 7: Data analysis and data evaluation with Excel
  - Frequency distribution, normalization and chi-square
  - Understanding the formulas by using intermediate steps

Data Analysis

Exercise 2: Calculating normalized figures

- Open the data file `data/distr_vfull_pos-reg_brown_matrix.txt` in LibreOffice/Excel
- Rename the sheet `rawfreq`
- Create a new sheet, rename it `fpm` (frequency per million) and copy paste the `rawfreq` table into this sheet
- We will calculate the normalized frequencies in this new table.
- Delete all figures from the `fpm` table
- Download the file with the BROWN corpus sizes and save it in the `data` directory of the course directory: brown_csizes_full.txt
- Choose `Tabelle -> Tabelle aus Datei einfügen` to open the BROWN corpus size file in a separate sheet
- Rename this sheet `corpsize`
- Now we can add our formula to the first cell in our table.
- Choose the first data cell in the `fpm` table (`VVD - A`)
- Write a `=` to indicate a formula
- Add the formula for normalization you may choose (select) the respective cells from the `rawfreq` and `corpsize` table (click and ENTER)
- The results of the formula will be displayed in the respective cell
- Add the formula to all cells

Data Analysis

- Session 8: Manipulating data sets with R
  - Basic notions related to R
    - Adding column names, adding columns, summarizing the data, merging data sets

Adding columns to a data frame

The data set does not yet contain meta information about the corpus it was extracted from. Thus, we add the columns

- `corpus` with the value `brown`
- `year` with the value `1961`
- `lgvar` (for language variety) with the value `AE`

If we would simply assign the values to R as follows, the variables would automatically be assinged the class `character`.

```
d.brown$corpus <- "brown"
d.brown$year <- "1961"
d.brown$lgvar <- "AE"
```

Data Analysis

- Session 9: Normalization and frequency distribution with R

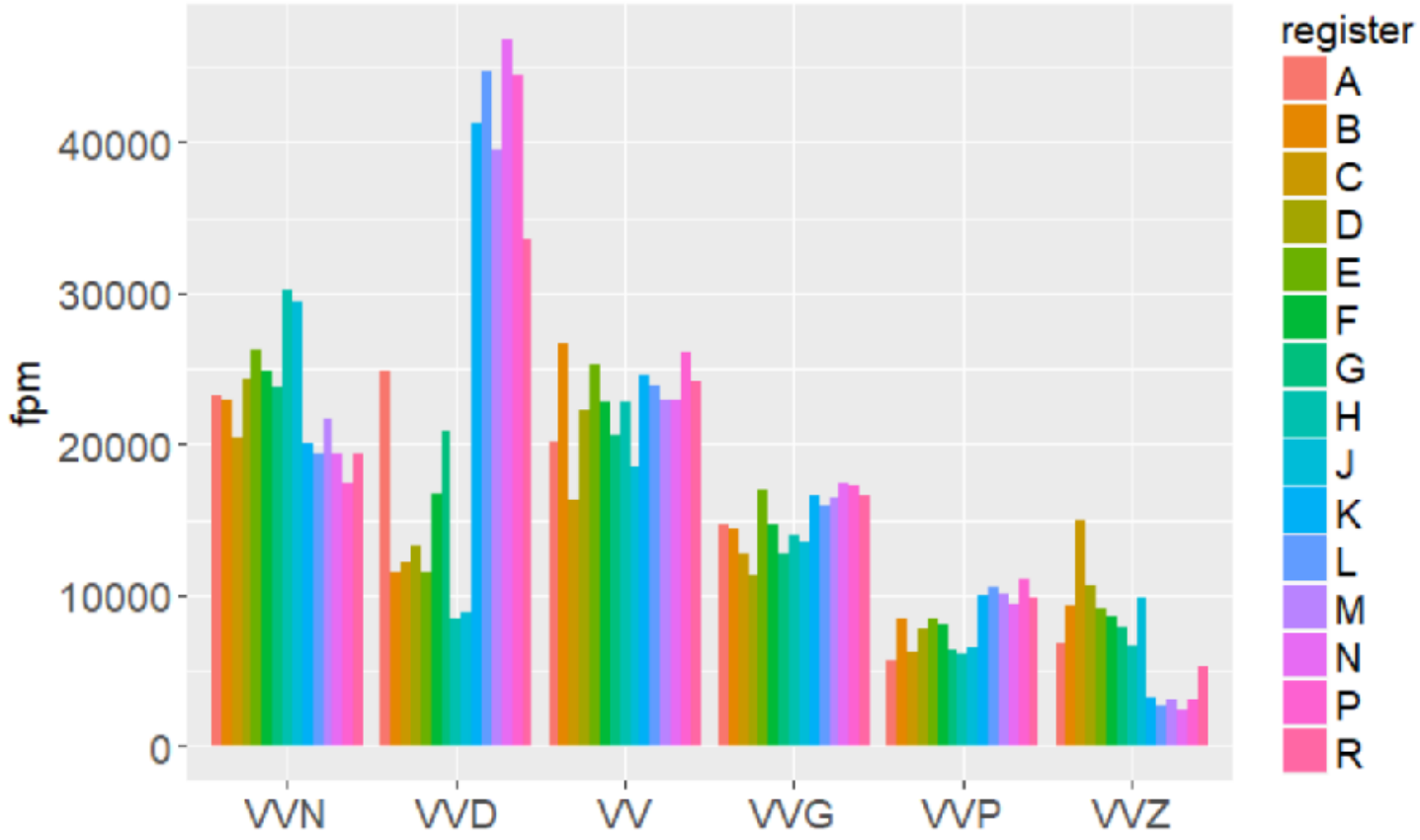
```
- setting parameters (`feat1` and `feat2`)
- normalize the data
- plot the data
```

```
```{r freq-distr-pos-reg}
# setting parameters
feat1 <- "pos"
feat2 <- "register"
# normalize data
d <- norm.data(dat,d.csize,feat1,feat2)
d$d
d$d.sum
# `d` is now a large list with 3 elements:
# d$d our tidy data set additionally including the subcorpus size
# d$d.sum a newly created data set including the counts for frequency and fpm
# d$l.feat.sorted a list of the values of `feat1` sorted after frequency
# fpm means frequency per million

# plots the summarized data for the specified subcorpus and linguistic feature
plot.bar(d$d.sum,feat1,"fpm",feat2)
```
```

Data Analysis

- Session 10: Plotting analysis results with R



Feedback from Students

"Meine Privatsphäre wurde durch die Überwachungskamera eingeschränkt."

"Die Vorlesung ist auf Englisch, aber die Übung ist auf Deutsch. Das kann ich manchmal daran nicht angewöhnen."

"Sehr schnell, manchmal überflogen."

"Manchmal ging es etwas schnell, aber da liegt es auch in der Verantwortung der Studierenden das anzumerken."

Feedback from Students

"Weiteres Eingehen auf kompliziertere Aufgaben"

"Jedes Thema wurde ausführlich besprochen & diskutiert, die Übungen waren sehr zielführend und gut gestaltet."

"Es wurde viel Wert darauf gelegt, dass alle mitkommen & der Stoff verstanden wird."

"Die überaus nette und freundliche Dozentin"

"Die Darstellung des Stoffes, sehr nette Professorin"

"Sehr nett und hilfreich"

"Die Dozentin hat viel gearbeitet und uns besonders motiviert."

"Eigene Arbeit am Computer mit Programm"

"Dass die Dozentin sehr hilfsbereit war und sich um alle Probleme gekümmert hat."

"Die Hilfestellungen, wenn etwas gar nicht funktionieren wollte und die Dozentin alles persönlich nochmal erklärte"

Summary

- Tutorials for
  - University courses
  - Self learning
- Reproducible sample study and exercises
  - Simulation of all steps of a “real” study
- Modular basic scripts
  - Reusable and adaptable to own future study
- Active and collaborative learning
  - Deeper understanding
  - Problems can be addressed and solved together immediately

Link to Website

http://fedora.clarin-d.uni-saarland.de/teaching/Corpus_Linguistics/