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

Mihaela Vela
Language Science and Technology
Saarland University
m.vela@mx.uni-saarland.de

Hannah Kermes
Language Science and Technology
Saarland University
h.kermes@mx.uni-saarland.de

Abstract

We present a practical course in corpus linguistics meant to provide students with a humanities background with the necessary knowledge and skills for an empirical study as basis for term papers, BA- or MA-thesis. The course is part of a new Bachelor program and is combined with a theoretically oriented course on corpus linguistics. The challenge is to provide students with the necessary understanding of the underlying concepts and skills of corpus linguistics without overwhelming them with too much technical detail. The course material is modular, allowing for easy updates, modifications and adaptations as well as reusable for different target groups, settings, and applications.

1 Introduction

In this paper we present a practical course in corpus linguistics, which is meant to provide students with a humanities background with the necessary knowledge and skills for an empirical study as basis for term papers, BA- or MA-thesis. The course is part of a new Bachelor program *Language Science* and is combined with a theoretically oriented course on corpus linguistics. The students in the program come from various backgrounds including translatology and language studies among others. Most of the students have only little or no experience in natural language processing.

The challenge is to provide them with the necessary understanding of the underlying concepts and skills of corpus linuistics without overwhelming them with too much technical detail. The course material is modular, allowing for easy updates, modifications and adaptations as well as reusable for different target groups, settings, and applications. The described processes, analysis and exercises are reproducible and portable to further studies.

In the following we will discuss challenges for teachers and students (Section 2) and describe the general concept of the course (Section 3) and its composition (Section 4). We conclude with a brief summary and envoy (Section 5).

2 Challenges for teachers and students

A practical course on corpus linguistics for students with a humanities background has challenges for both teachers and students.

The challenges stem from the seemingly opposed character of the digital applications and the humanities disciplines as well as from the character of a practical course requiring a lot of active learning on the side of the students.

Challenges for teachers include:

- motivating students and lowering the psychological and practical barriers
- trying to avoid or solve technical problems
- dealing with heterogeneous groups both with regard to the prior knowledge of the students as well as with their different learning pace
- keeping track of the learning success of the group and individual students, adjusting the teaching speed and/or type accordingly

Challenges for students include:

- engaging with a potentially new kind of subject matter
- dealing with and solving technical problems
- coping with the high demands of active learning

Motivating students to engage with the technical aspects of corpus linguistics often boils down to

answering the question about the usefulness of the methodology. Good and obvious examples for applications in the students' discipline(s) exemplify the additional value. Useful to this respect are simple and understandable practical exercises to exemplify and to help lower potential psychological barriers with regard to technical applications. Motivating students stays a challenge throughout a course as technical aspects can easily become cumbersome and tedious. Active learning plays an important role in this respect. Active learning in the sense of an instructional method engaging students in meaningful learning activities in the classroom (e.g. doing exercises, working on and discussing problems/results) (Prince, 2004; Bonwell and Eison, 1991). It allows to keep the students active and involved giving them an immediate feedback on their learning success. The broader goal of the session, however, should be made clear to show the necessity of the activities of the students.

Technical problems with regard to applications and code are a challenge for both teachers and students. A lot of technical problems, e.g. with installing software, can be avoided by using online tools, e.g. online corpora or web services for corpus annotation. Another possibility to limit technical problems is to provide sample code for more complex examples, which only needs to be modified or complemented for exercises or later application. This can help to focus on the main aspects of the methodology as the technical difficulties are reduced to a minimum. Nevertheless, technical problems cannot be avoided completely, it can even be good to provoke particular problems in class. Problem solving, especially finding error in code or regular expressions, is also an aspect of corpus linguistic research. Working through examples and exercises in class will inevitably lead to some technical problems. However, as they occur in class, the teacher can immediately provide help in finding and solving the problems.

Teachers are often confronted with heterogeneous groups, both with regard to their prior knowledge as well as with different learning speeds. Although this is a general challenge in teaching, the groups are often more heterogeneous in digital humanities settings and are especially pronounced when teaching technical skills. Again an active learning environment where examples and exercises are worked through in class can help to adopt to individual needs. It is easier for the teacher to find out about individual problems and to provide immediate support. It is also possible to provide exercises for different levels or extra exercises for more advanced students. Working on a problem as a group can also foster deeper understanding.

Making students present the results of exercises and discussing them as a group helps teachers keeping track of the learning success of the group and of individual students. This is important to eventually adjust the teaching speed and/or type accordingly. The discussion of the results can also be used to sum up and point to important aspects of a teaching unit. This helps the students to reflect on their own learning success and to evaluate the personal organization and structure of their learning activities.

In the following we will now describe the general concept of the course and how it addresses the challenges described above.

3 General concept

The course covers the two main aspects of corpus linguistics: (i) corpus building and (ii) corpus analysis. The main emphasis, however, is on corpus analysis as depicted in Figure 1. Tutorials lead the students through the main steps in corpus building from the digitized text to an annotated searchable corpus and from the linguistic research question and corpus extraction to corpus analysis.

The course is constructed like a sample study, each tutorial representing a particular step in the process. In this sense, the tutorials in both parts build on one another, as each tutorial produces the input data for the next. However, as the necessary sample data is provided at the beginning of each tutorial, the tutorials are also self-contained. In the first part, we create a corpus out of a small plain text sample, adding meta data and basic linguistic annotation. In the second part, we look at a sample research question extracting and analyzing the respective data. The characteristic of the course of a sample study allows the students to get acquainted with the process of an empirical corpus linguistic study, facilitating a later application of the methodology to a study of their own.

The course material is provided as a website with detailed step-by-step tutorials including the necessary background information, examples with sample data and exercises. Links to external knowledge sources and tutorials, provide access to additional information including more (technical) details, more profound background or more complex

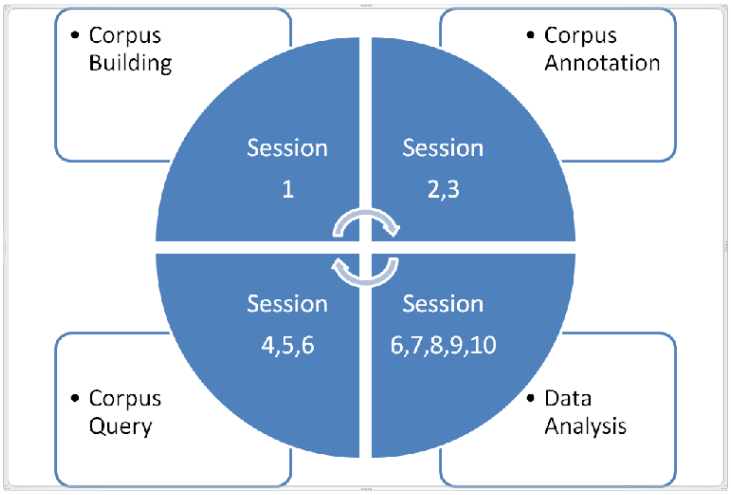

Figure 1: Structure of the course

applications. The tutorials may be worked through at individual pace adapting to the specific needs of different target groups and individual students, skipping sections or providing additional information or exercises.

The tutorials are written in `R-Markdown` and converted into HTML websites. Using `R-Markdown` as source documents has several advantages: (i) the tutorials are easy to modify, (ii) additional information as well as new material can easily be integrated, (iii) the students can download the source document, which allows for individual notes as well as to reproduce the provided sample analysis.

Although the tutorials were initiated as course material for a university course, they can also function as self-learning tutorials or as a (knowledge) base and memory hook for later use when writing a term paper, BA- or MA-thesis. As a university course it follows the concept of inverted or flipped classroom (Bergmann and Sams, 2012; Handke et al., 2012) in the sense that sample study and exercises are worked through in the classroom individually or as a group, followed by a group discussion. This allows to address problems immediately, discuss them as a group, work through advance concept and engage in collaborative learning and problem solving (Tucker, 2012) again more or less simulating a research process in a team.

4 Tutorial Description

The practical course described here consisted of ten incremental sessions on Corpus Linguistics. The goal, as described in the previous section, is to get students acquainted with the concepts and notions of Corpus Linguistics. Each session was planned as an interactive combination between a tutorial (presented by the lecturers or prepared by the students) and the corresponding exercises (solved in class by the students with the lecturer's assistance). The tutorial corresponds to the theoretical part, introducing a new topic, while the exercises are meant as practical applications.

In this section we describe the course in detail including not only the structure, content and technical aspects of the course, but also the necessary infrastructure to teach this type of classes. We decided to publish the entire material of the course online[1], facilitating the availability and reproducibility of all course materials. As mentioned before, we use `R Markdown` for this purpose, being able to combine both narrative text and code and producing formatted output.

As shown in Figure 1, the course is structured into four blocks, *Corpus Building*, *Corpus Annotation*, *Corpus Query* and *Data Analysis*, distributed over ten sessions. Each new session introduces a new concept, but at the same time using previously introduced concepts. In terms of corpora we use the Royal Society Corpus (RSC) (Kermes et al., 2016), a historical corpus of written scientific English, as well as the BROWN (Francis and Kučera, 1979), FLOB (Mair, 1999b), LOB (Johansson and Goodluck, 1978) or FROWN (Mair, 1999a) corpora, covering different time periods and registers

[1] `http://fedora.clarin-d.uni-saarland.de/teaching/Corpus_Linguistics/index.html`

for both American and British English.

The course is structured as follows.

- Session 1: Corpus building with XML and TEI
- Session 2: Tagging with TreeTagger
- Session 3: Corpus annotation with WebLicht
- Session 4: Corpus query with regular expressions
- Session 5: Corpus query with patterns
- Session 6: Data extraction and data formats
- Session 7: Data analysis and data evaluation with Excel
- Session 8: Manipulating data sets with `R`
- Session 9: Normalization and frequency distribution with `R`
- Session 10: Plotting analysis results with `R`

Session 1 belongs to the *Corpus Building* block and provides an introduction to XML (EXtensible Markup Language) and TEI (Text Encoding Initiative). The goal of this class is to make students understand the importance and the syntax of mark-up languages when working with corpora. The first class starts with an exercise by asking students to mark title, paragraphs and sentences in a .txt file. The different solutions are meant to show the possible variation in marking linguistic units, here title, paragraphs and sentences, and making a point why a standardized mark-up language is necessary. The tutorial introduces first the XML syntax, followed by the TEI syntax. For completion the session ends with an exercise on encoding the same text according to the TEI guidelines.

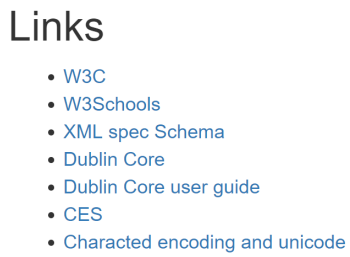

Figure 2: Additional resources for data encoding.

As described in Section 3, we provided additional information, beyond the scope of the specific class. In Session 1 this additional information was provided as links depicted in the Figure 2 below.

Session 2 and Session 3 are part of the *Corpus Annotation* block. Session 2 deals with part-of-speech tagging, including its definition and the introduction of the concept of a tagset. More specifically, this class provides also an introduction to the usage and configuration options of the TreeTagger (Schmid, 1994; Schmid, 1995). The exercise deals with the installation and usage of the TreeTagger as well as performing tagging on .txt files, but also on .xml files, as depicted in Figure 3.

Session 3 goes one step further by introducing additional annotation layers to the already existing part-of-speech annotation. This is carried out by WebLicht (Hinrichs et al., 2010), a web based environment for the annotation of corpora. It includes tools for tokenization, lemmatization, pos-tagging and parsing (among others), which can be combined individually to tool chains. The WebLicht tutorial describes the usage of the tool by depicting screenshots and giving examples. The exercise for the students is to build a processing chain in WebLicht including at least a tokenizer and the TreeTagger. The files used for this exercise are the same as in Session 2.

Session 4 and Session 5 are concerned with corpus query belonging to the *Corpus Query* block, being concerned with the qualitative analysis of the texts. Session 4 is meant as an introduction to regular expressions, defining the concept of a regular expression, but also explaining, by examples, the special characters and their role in formulating a regular expression. After practicing formulating queries with regular expressions in Notepad++[2] the students were introduced to the Saarland University CQPWeb (Hardie, 2012) platform and to the CQP syntax (Evert and Hardie, 2011)[3]. The tutorial consists of a series of examples for queries meant to consolidate the knowledge about regular expressions. The corresponding exercises consist of a set of queries to be carried out in CQPWeb on the RSC and BROWN corpora. An example of such an exercise can be found in Figure 4.

Session 5 is a continuation of Session 4 building on regular expression syntax extending the simple search queries introduced before to more complex queries including patterns. In the exercise part

[2] `https://notepad-plus-plus.org/`

[3] `https://corpora.clarin-d.uni-saarland.de/cqpweb/`

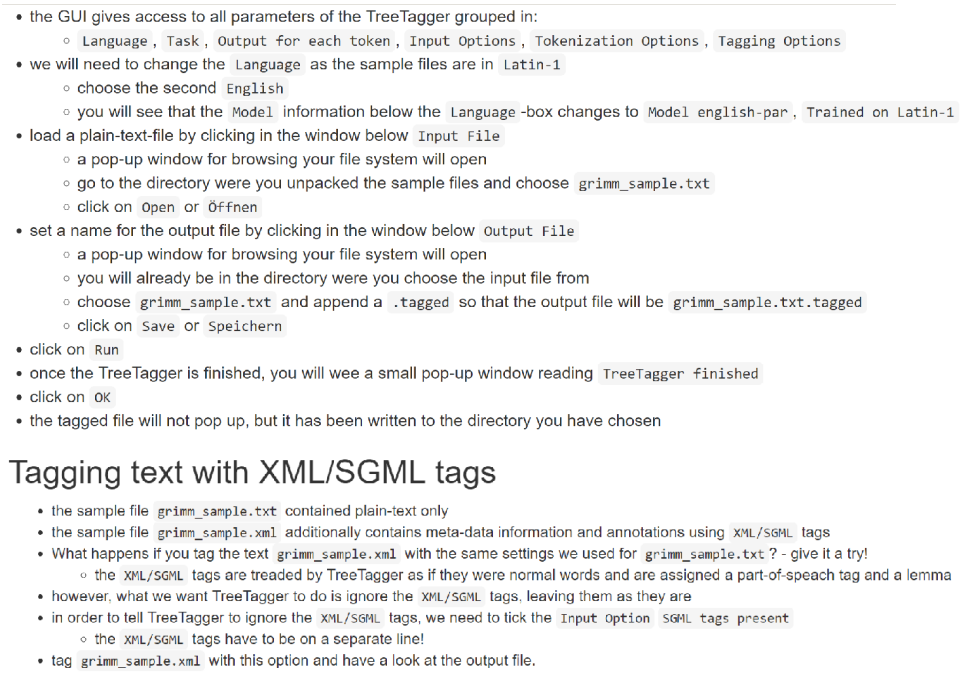

Figure 3: Instructions for tagging with the TreeTagger.

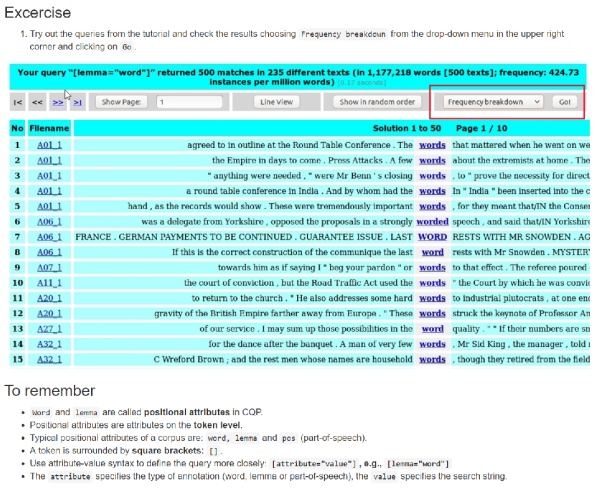

Figure 4: Introductory exercise in CQPWeb.

of the class the students are being asked to build their own pattern by using the CQP syntax and to query again the RSC and BROWN corpora using CQPWeb as shown in Figure 5.

Session 6 belongs to both the *Corpus Query* as well as the *Data Analysis* block, dealing with the results of a query. The result of a specific linguistic induced query is usually a data set containing information about a particular linguistic phenomenon extracted from a particular corpus. In this class students were provided with the concept of a data set, creating, formatting and manipulating it by using the online search tool CQPweb. Strongly related to the data set this class is introducing the notions of observations, features and values of features. During the practical part students were asked to create their own data set based on a research question (e.g. distribution of content verbs and their parts-of-speech across registers in a specific corpus) and to formulate the research question in terms of query, observations and features.

Session 7 to 10 belong to the *Data Analysis* block, introducing basic data analysis and data evaluation methods such as frequency distribution, normalization and statistical significance test using the χ^2 (chi-square) test. The statistical analyses in these sessions are based on the queries and data extracted in the previous sessions.

Session 7 is a gentle introduction to data analysis, introducing (with relevant examples) all theoretical notions related to frequency distribution, normalization and χ^2. The practical application of these concepts is realized by exercises executed in Excel/Libre Office. Excel/Libre Office is not the state-of-the-art in statistical analysis but has big advantage: the statistical analysis can be performed step by step permitting students to understand the path to the final formula. Understanding how a specific formula works (including the intermediate steps) is a great benefit for learners, who need to use this kind of knowledge later in their academic

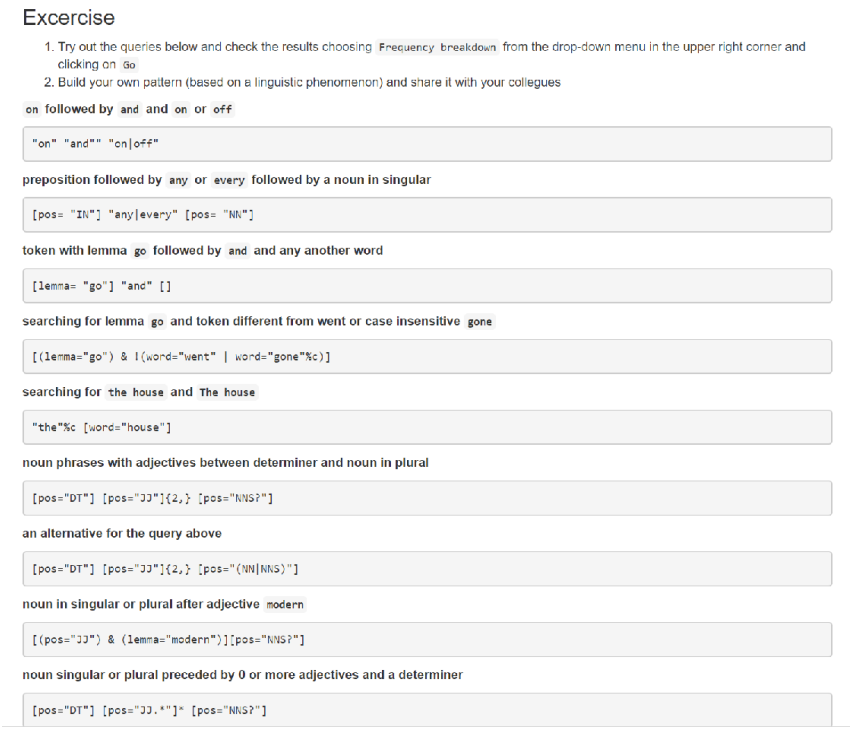

Figure 5: Exercises with patterns in CQPWeb.

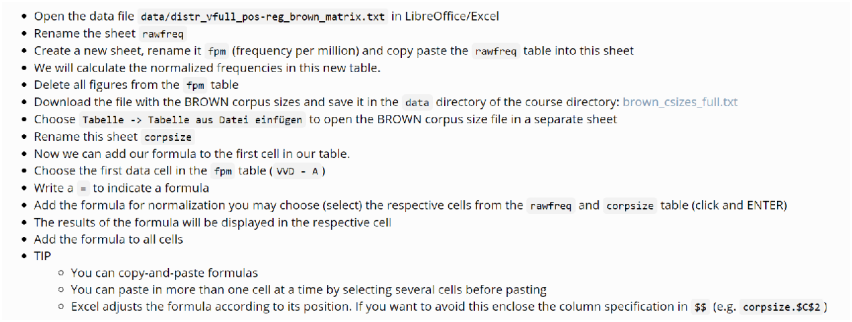

Figure 6: Calculating normalized figures in Excel/Libre Office.

studies.

Session 8 introduces statistical analysis with R in R Studio. It introduces basic notions related to R, including data manipulation such as adding column names, adding additional variables (columns), summarizing the data, merging and combining two or more data sets by presenting appropriate examples for each of these topics. In the exercise part of this session students are asked to extract similar data sets from other corpora applying the same data manipulation as used in the examples. At the end all data sets are combined to one large data set, which will then be analyzed in Session 9. Figure 7 shows the introductory part related to data frames in R.

Session 9 relates to Session 7 in that it is concerned with basic data analysis and data evaluation methods such as frequency distribution (see Figure 8), normalization and statistical significance test using the χ^2 test. The two sessions differ by the tools used for the analysis. While data analysis in Session 7 was exemplified using Excel, Session 9 uses R. The exercises from Session 7 are repeated in Session 9 to show the relation of the tools. However, it is also shown that R is more powerful when dealing with multivariate data sets extending the analysis performed in Session 7.

Session 10 is the continuation of Session 9 introducing additional aspects of data analysis showing how to visualize and interprete data from different perspectives (see Figure 8). Students prepare the different visualizations and their interpretation in

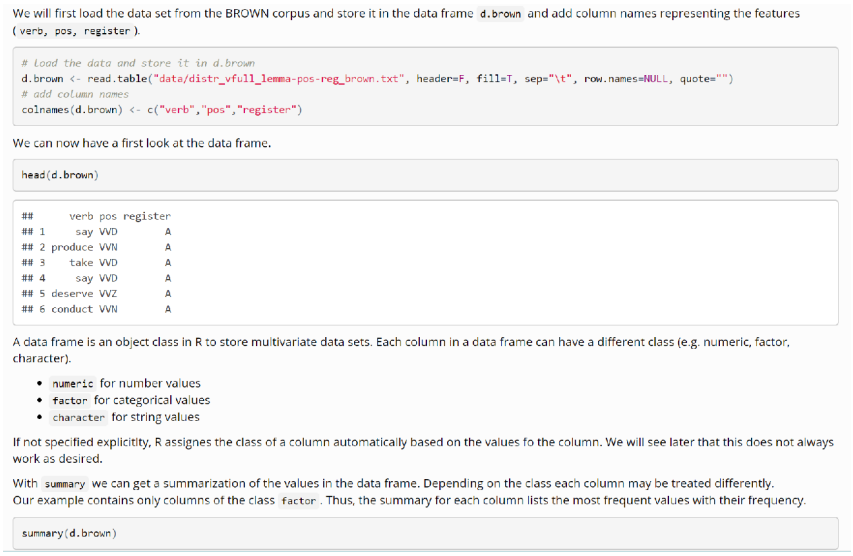

Figure 7: Exercises with data frames in R.

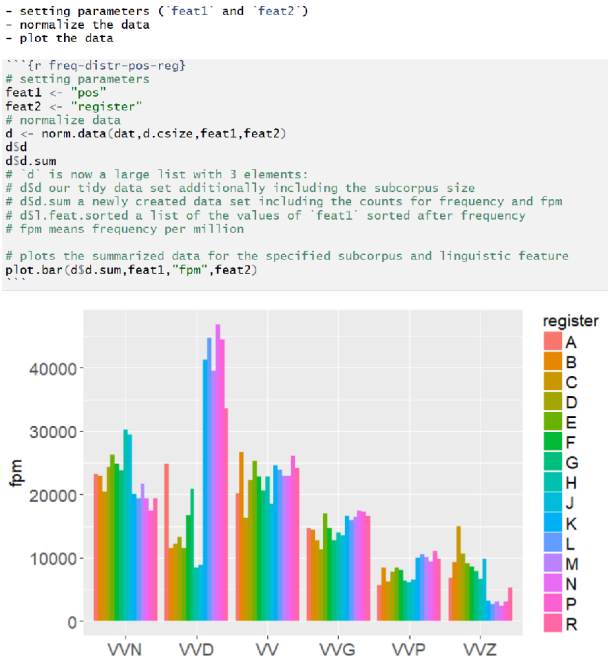

Figure 8: Data analysis in R.

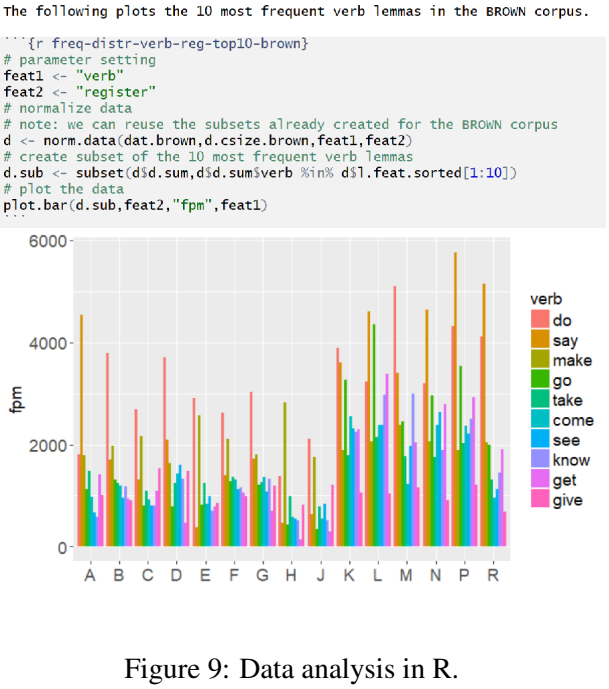

Figure 9: Data analysis in R.

groups. The results are then discussed together in class. The importance of verifying the interpretation of the macro perspective of the visualization with the micro perspective, the examples from the corpora (concordance lines) is made explicit in this discussions, linking and intertwining quantitative and qualitative analysis.

The `R Markdown` documents used throughout Session 8 to 10 include sample code for data manipulation and data analysis (see Figures 8 and 9). The documents are modular allowing to apply them to different data sets as well as copying, modifying and adapting the included code. The modification and adaptation is exemplified in the exercises and the students are encouraged to make notes about technical aspects and interpretations.

5 Conclusion

We presented a ten session practical course on Corpus Linguistics for students with a humanities background. The structure of the course is based on active learning methods to address the challenges

of teaching a technical course to students with little or no technical background. An active learning environment encourages students to work on research question alone or as a group, addressing (technical) challenges, solving technical and research related problems and discussing results. The role of the teacher moves in the direction of an assistant, answering questions, pushing in the right direction and helping to find solutions. The course and the course material, as presented here, allows for an easy modification, adaptation and extension of the course material. This makes the course and its material applicable to different target groups and settings, making the creation of such material worth the effort.

References

Jonathan Bergmann and Aaron Sams. 2012. *Flip Your Classroom: Reach Every Student in Every Class Every Day*. International Society for Technology in Education, Eugene, Or.

Charles C. Bonwell and James A. Eison. 1991. *Active Learning: Creating Excitement in the Classroom*. Number 1, 1991 in ASHE-ERIC higher education report. School of Education and Human Development, George Washington University, Washington, DC.

Stefan Evert and Andrew Hardie. 2011. Twenty-First Century Corpus Workbench: Updating a Query Architecture for the New Millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*, Brimingham, UK.

W.N. Francis and H. Kučera. 1979. *Manual of Information to Accompany A Standard Corpus of Present-day Edited American English, for Use with Digital Computers*. Brown University, Department of Lingustics.

Jürgen Handke, Alexander Sperl, and Deutsche ICM-Konferenz, editors. 2012. *Das inverted classroom model: Begleitband zur ersten deutschen ICM-Konferenz*. Oldenbourg, München. OCLC: 810266426.

Andrew Hardie. 2012. CQPweb –Combining Power, Flexibility and Usability in a Corpus Analysis Tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

Erhard Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. WebLicht: Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29. Association for Computational Linguistics.

Geoffrey Leech Johansson, Stig and Helen Goodluck, 1978. *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*.

Hannah Kermes, Jörg Knappen, Stefania Degaetano-Ortlieb, and Elke Teich. 2016. The royal society corpus: From uncharted data to corpus. In *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'16)*, Portoroz, Slovenia.

Christian Mair, 1999a. *The Freiburg-Brown Corpus (Frown)*.

Christian Mair, 1999b. *The Freiburg-LOB Corpus (F-LOB)*.

Michael Prince. 2004. Does active learning work? A review of the research. *Journal of engineering education*, 93(3):223–231.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

Bill Tucker. 2012. The flipped classroom. *Education next*, 12(1).