Government of Russian Federation

Federal State Autonomous Educational Institution of High Education
«National Research University Higher School of Economics»

School of Linguistics

Syllabus for the course

Linguistic data: quantitative analysis and visualisation

Master’s program “**Linguistic Theory and Language Description**”**,** 45.04.03 “Fundamental and Applied Linguistics”, Master level

Authors of the course:

Olga Lyashevskaya, HSE, School of Linguistics, Professor, olyashevskaya@hse.ru
George Moroz, HSE, School of Linguistics, Senior Lecturer, agricolamz@hse.ru
Ilya Schurov, HSE, Department of Higher Mathematics, Associate Professor, ischurov@hse.ru
Alla Tambovtseva, HSE, Department of Higher Mathematics, Assistant, atambovtseva@hse.ru

Approved by the meeting of the School of Linguistics, Faculty of Humanities on 05 June, 2018
Head of the School of Linguistics E. V. Rakhilina
Recommended by the meeting of the MA program Academic Council resolution № 2, on 05 June 2018
Approved by the head of the MA program Academic R.N. Krivko on 05 June 2018

National Research University Higher School of Economics
Syllabus for the course "Linguistic data: quantitative analysis and visualisation" for the programme 45.04.03 "Fundamental and Applied Linguistics", Master level

Moscow, 2018

This syllabus cannot be used by other university departments and other higher education institutions without the explicit permission of the School of Linguistics

1. Course Description

This course is an introduction to key quantitative approaches to the analysis of linguistic data.

Pre-requisites

The course is based on the student's knowledge of fundamentals of empirically-based linguistic analysis. Knowledge of fundamentals of probability and statistics as well as previous experience in programming and machine learning is considered a plus but not required. All teaching is conducted in English.

Course Type (compulsory, elective, optional)

The course is part of the compulsory courses offered by MA program "Linguistic Theory and Language Description" and it is elective for students of MA program "Computer linguistics". It is a blended course which combines in-class lectures and seminars, self-running sessions which include individual and group work, and attending an open online course using the Datacamp platform.

2. Learning Objectives

Within this course you will:

- learn about the principal steps of a quantitative research in linguistics;
- learn about the possibilities and limitations of quantitative approaches as applied to different research questions;
- learn to formulate research questions and develop them into testable hypotheses;
- explore the possibilities of data collection and different approaches to sampling;
- learn to evaluate the quality of a quantitative approach;
- study the most common corpus, experimental, and mixed design of the linguistic studies and learn to evaluate research plans, discover and prevent the associated threats to data validity;
- practice in preparing your quantitative data for analysis, evaluating the quality of your data; treating missing data;
- learn about the possibilities and limitations of conventional statistical techniques and criteria, as well as some popular contemporary multivariate statistical methods;
- learn to choose and apply in practice a set of appropriate statistical tests for

your research question.

3. Learning Outcomes

On completion of the course, you will be able to:

- account for basic types of data used in linguistic research;
- apply basic quantitative methods for analysing linguistic data;
- critically discuss the limitations of commonly used methods for answering research questions about language;
- reason on how to interpret linguistic results, including how to evaluate what kind of information a given method can offer and how to estimate the potential range of variables that can affect results in linguistic research;
- critically evaluate linguistic data presented in previous research;
- apply different techniques for presenting both qualitative and quantitative linguistic data in scholarly writing.

4. Course Plan

1. Introduction to R. Types of data. Dataframe. Functions and arguments.
2. Descriptive statistics. Basic visualizations.
3. Dplyr style in R, pipes. Visualizing data with ggplot2.
4. Hypothesis testing. Types of distribution. P-values. Exact binomial test, t-test, ANOVA. Confidence intervals. Chi-squared and Fisher exact test.
5. Correlation
6. Regressions: linear and polynomial
7. Logistic regression
8. Fixed and random effects. Mixed-effects models
9. Bootstrap. Decision trees. Decision forests
10. Distance matrices. Clusterization.
11. Dimension reduction, visualisations using MDS, PCA, CA, MCA.
12. Bayesian statistics

5. Reading list

Compulsory reading:

1. Levshina N. How to do Linguistics with R. Amsterdam: John Benjamins Publishing Company, 2015. ISBN: 978-90-272-1224-5, 978-90-272-6845-7. [eBook available in the university library]

2. Gries St. Th. Statistics for Linguistics with R: A Practical Introduction. De Gruyter Mouton, 2013. ISBN: 978-3-11-030728-3, 978-1-299-72482-2, 978-3-11-030747-4. [eBook available in the university library]

Optional reading:

3. Hadley W. Ggplot2: Elegant Graphics for Data Analysis. Penn. Springer. ISBN: 978-3-319-24275-0, 978-3-319-24277-4. [eBook available in the university library]

4. Desagulier A. Corpus Linguistics and Statistics with R. Springer eBooks, 2017. ISBN: 978-3-319-64570-4, 978-3-319-64572-8. [eBook available in the university library]

6. Grading System

The course is examined through continuous assessment of written assignments and the final project.

Written assignments includes theoretical tests and practical problem-solving. The assignments are published online. The assignments should be submitted via an electronic form. The submission after the deadline will lead to penalty: 10% for delay within 1 hour, 20% for delay within 1 week, 50% for delay within 1 month, 90% for delay for more than 1 month. The grade of every written assignment is a floating point number from 0 to 10. The average of all written assignments (with equal weights) is student's Homework Score.

The student is expected to prepare the final project in a written form as electronic document.

The colloquium is conducted as a discussion of the final projects. The exam is conducted in the form of oral defense of the final project. The Exam Score measures the overall quality of the final project. It is integer number from 0 to 10.

The Cumulative Score is obtained from the following formula:

Cumulative Score = 0.8 × Homework Score + 0.2 × Colloquium Score

The Final Score is obtained from the following formula:

Final Score = 0.6 × (Cumulative Score) + 0.4 × (Exam Score).

7. Guidelines for Knowledge Assessment

Examples of homework assignment problems:

1. Plot a histogram for a given sample.
2. What kind of statistical test should be used to answer the question: can we claim that the mean length of sentence in texts of author A is larger than in texts of author B? What kind of data needed to apply this test? Use given data to answer the question.

3. Find a regression coefficients for given data and statistical model. Interpret them in terms of domain-specific problem
4. Visualise given multidimensional data using dimension-reduction techniques (e.g. PCA or MDS).
5. Apply hierarchical clustering to a dataframe containing properties of different languages. Plot and interpret dendrogram.

Final project should include the following parts:

- Research objectives and hypothesis to be tested.
- Description of input data.
- Discussion of the methods of analysis and their applicability.
- Obtained results and their interpretation.
- All the code used.

8. Method of instructions

Lectures are devoted to theoretical foundations of quantitative methods, including statistics. Practical sessions will review concepts taught in lectures and connect them to homework problems through examples. In-class sessions will provide hands-on-experience opportunity: the students will be asked to follow the presenter using their laptops and they will work on assigned exercises while in session. Online courses will be used to conduct some of the homework assignments and review sessions.

9. Special Equipment and Software Support

We will primarily rely on R free software to demonstrate and operationalize concepts presented during lectures. Students are expected to download and install free R software (https://www.r-project.org) or use rstudio.cloud online environment (https://rstudio.cloud).

As online course, we will use DataCamp free course "Introduction to R" https://www.datacamp.com/courses/free-introduction-to-r.