Government of Russian Federation

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

School of Linguistics
Master's programme 'Linguistic Theory and Language Description'

# Syllabus for the course "Linguistic data: quantitative analysis and visualisation"

Author:
O. Lyashevskaya, olyashevskaya@hse.ru
G. Moroz, agricolamz@hse.ru
I. Schurov ischurov@hse.ru

Approved by the meeting of the Master Program Academic Council
(day/month/year) 30.05. 2016

Moscow, 2016

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

## 1. Scope of Use

This course is an introduction to key quantitative approaches to the analysis of linguistic data. The course is taught in the form of lectures, seminars and self-running sessions which include individual and group work and attending open online courses. All teaching is conducted in English.

## 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, the student 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. Role of the course within the structure of Master program

The course aims to provide students with knowledge and competencies necessary to plan and conduct research projects of their own leading to M.Sc. dissertation and scientific publications.

This course is required as part of two M.Sc. programs, “Linguistic Theory and Language Description”, and “Computational Linguistics”.

## 5. Schedule

1. Introduction to R. Types of data. Dataframe. User's functions
    a. 1 lecture
    b. 3 seminars
        i. Base R objects. Simple plots. Simple functions.
        ii. Dataframes. Read and write data. Manipulation with data (R base vs. dplyr)
        iii. Base R functions code. User’s functions. Loops. Functions in functions. Default arguments. Optional arguments.
2. Visualization.
    a. 1 lecture
    b. 2 seminars
        i. Base R vs. ggplot2: 2 variables, 3 variables, many variables
        ii. RMarkdown. Tables in RMarkdown. Mixing different languages in RMarkdown
3. Descriptive statistics. Confidence intervals. T-tests. P-values. Normality tests
    a. 1 lecture
    b. 3 seminars
        i. Descriptive statistics, boxplot, violinplot. NAs.
        ii. T-tests. P-value.
        iii. Z-score. Confidence intervals. CI vs. p-values
4. $\chi^2$. Fisher exact test
    a. 1 lecture
    b. 1 seminar
5. Correlation
    a. 1 lecture

   b. 1 seminar
6. Regressions: linear and polynomial
   a. 1 lecture
   b. 1 seminar
7. ANOVA
   a. 1 lecture
   b. 1 seminar
8. Logistic regressions
   a. 1 lecture
   b. 1 seminar
9. Mixed-effects models
   a. 1 lecture
   b. 1 seminar
10. Bootstrap. Decision trees. Decision forests
    a. 1 lecture
    b. 2 seminars
       i. Confidence intervals again. Bootstraping.
       ii. Decision trees. Decision forests.
11. Clusterization
    a. 1 lecture
    b. 2 seminars
       i. Different types of distance matrices. K-means
       ii. Hierarchical clusterization. DBSCAN.
12. MDS, PCA, CA, MCA
    a. 1 lecture
    b. 2 seminars

## 6. Requirements

The student is expected to

- be able to formulate the research problem in formal terms;
- know all the relevant notions;
- understand the theoretical background of methods discussed as well as their limitations;
- be able to use the software to process data;
- be able to give correct interpretation of the output of the software in terms of the research problem.

## 7. Assessment

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 deadline for each assignment is specified upon publishing and will never be postponed.
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) rounded to integer with Google Spreadsheet's ROUND function is student's Cumulative Score.

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

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 Final Score is obtained from the following formula:

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

Rounding with Google Spreadsheet's ROUND function is applied.

## 8. Course Description

Introduction to R. Types of data. Dataframe. User's functions.

Basic R objects. Simple plots. Simple functions. Dataframes. Converting data. Reading and writing data. Attaching data. Working with Unicode and other encodings.

Manipulation with data (R base vs. dplyr). Basic R functions code. User's functions. Loops. Functions in functions. Default arguments. Optional arguments.

R Studio.

Visualization.

Base R vs. ggplot2: two variables, three variables, many variables.

RMarkdown. Tables in RMarkdown. Mixing different languages in RMarkdown.

*Literature*:

Levshina 2015: 69-86 (Chapter 4). Tufte 2001. Murrell 2005. Wickham 2009

Descriptive statistics. T-tests. P-values. Normality tests. Confidence intervals.

Descriptive statistics, boxplot, violinplot. NAs.

Null and alternative hypotheses. Statistical significance. Significance level.

T-tests. P-value. Misconceptions about p-value. One-tailed and two-tailed tests. Wilcoxon and Mann-Whitney tests.

Z-score. Confidence intervals. Standard error. CI vs. p-values.

*Literature*:

Levshina 2015: 41-68 (Chapter 3). Levshina 2015: 87-114 (Chapter 5).

$\chi^2$-test. Fisher's exact test.

Contingency tables. Observed and expected frequencies. The χ2 -test of independence. Mosaic plots. χ2 -test and big data. Effect size metrics. Fisher's exact test.

*Literature*:

Levshina 2015: 199-222 (Chapter 9).

Correlation

Relationship between two quantitative variables. The Pearson product-moment correlation coefficient. Correlation vs. paired t-test. Spearman's *rho* and Kendall's *tau* rank correlation. Correlograms.

*Literature*:

Levshina 2015: 115-138 (Chapter 6).

Regressions: linear and polynomial

Linear regression with several explanatory variables. Organizing data and making hypotheses. Selecting the explanatory variables. Checking for overfitting. Linear regression with categorical explanatory variables. Polynomial regressions.

*Literature*:

Levshina 2015: 139-170 (Chapter 7).

ANOVA

Analysis of variance (ANOVA): finding differences between several groups. Reporting results of ANOVA. Repeated-measures and mixed ANOVA. Other types of ANOVA: analysis of covariance (ANCOVA), multivariate analysis of variance (MANOVA).

*Literature*:

Levshina 2015: 171-198 (Chapter 8).

Logistic regressions

Binomial logistic regression modeling. Logistic function. Intercept and slope. Checking for overfitting. Polytomous logistic regression modeling.

*Literature*:

Levshina 2015: 139-170 (Chapter 7).

Mixed-effects models

Fixed and random effects. Research design with random-effect variables. Random intercepts. Random slopes. Linear mixed-effect modeling. Logistic mixed-effect modeling.

*Literature*:

Baayen 2008: 241-302.

Levshina 2015: 192-196.

Baayen et al. 2013: Making choices in Russian: pros and cons of statistical methods for rival forms. Russian Linguistics 37: 253-291.

Bootstrap. Decision trees.  Decision forests.

Bootstrapping. When bootstrap?

Classification and regression trees. Bias and overfitting. Random forests. Importance of explanatory variables.

*Literature*:

Levshina 2015: 291-300 (Chapter 14).

Clusterization

Distances in a multidimensional space. Distance metrics (Euclidean, Manhattan, Maximum). Bottom-up and top-down approach (agglomerative and divisive clustering). Methods of clustering (complete, single/nearest neighbors, average, ward). Which number of clusters is optimal? Validation of cluster solution. AU/BP values.

*Literature*:

Levshina 2015: 301-321 (Chapter 15).

MDS, PCA, CA, MCA

Multidimensional Scaling.

Principal Component Analysis. Contributions of components of PCA to explaining variance. Variables factor map. Individuals factor map.

Simple and multiple correspondence analysis for categorical variables.

*Literature*:

Levshina 2015: 333-350 (Chapter 17). Levshina 2015: 353-361 (Chapter 18). Levshina 2015: 367-385 (Chapter 19).

## 9. Bibliography

1. Levshina, Natalia. 2015. How to do Linguistics with R. Info Companion web site

2. Gries, Stefan Th. 2009. Quantitative Corpus Linguistics with R. Info Companion web-page
   2nd edition, 2013. Главы Hierarchical Cluster Analysis, Mixed Effects models.
3. Baayen, Harald. 2008. Analyzing Linguistic Data: A Practical Introduction to Statistics using R. Google books
4. Oakes, Michael. 1998. Statistics for Corpus Linguistics. info
5. Johnson, Keith. 2008. Quantitative Methods In Linguistics. info
6. Cantos Gómez P. 2013. Statistical methods in language and linguistic research. Sheffield; Bristol. info
7. Butler, Christopher S. 1995. Statistics in Linguistics. Web-edition
8. Tufte E. R. 2001 The Visual Display of Quantitative Information. info
9. Wickham H. 2009 ggplot2. Elegant Graphics for Data Analysis. Info
10. Murrell P. 2005 R Graphics Companion web-page