Final projects

March 23, 2022

Linguistic Data: Quantitative Analysis and Visualisation

Linguistic theory group

Course webpage

In your final project, you have to conduct empirical linguistic research using methods discussed during the course. Project consists of the following steps:

1. Choose and state research question and hypotheses you want to test.
2. Propose research design. Specify, what kind of data you need to test your hypothesis, how are you going to collect this data and analyse it.
3. Collect the data.
4. Use statistical methods to analyse the data.
5. Interpret results of your statistical analysis in terms of the original research question and hypotheses.
6. Prepare the report that should include the description of all previous steps.
7. Defend your research during short oral presentation.

You can prepare your project individually or in pair with other student of the program. In the latter case, you should indicate personal contribution of both authors into the project, and both authors have to be able to answer questions related to any part of the project.

Step 1. Research questions and hypotheses

You are free to choose any research question in the field of linguistics you are interested in that can be studied empirically with quantitative methods. This research can be part of your ongoing term work or other ongoing research projects

you participate in (i.e. if you work in some of HSE Labs), but it should not repeat research you've done before, except possible data collection step. If in doubt, consult with your instructors.

Choice of the research question is usually connected with the development of research design: you should state question that can be answered with the data you can collect and methods you know how to apply. As an example, you can conduct corpus-based research and extract data from some available corpora using corresponding corpus methods. Another option is to do your own linguistic experiments, surveys, interviews, etc.

For any hypothesis you stated, it is a good idea to provide some justification why do you believe this hypothesis is plausible.

If you need examples of possible types of topics for your projects, you can consult projects done in previous years (of different quality, for orientation only).

If you don't have any ideas on the research topic or need any other help, do not hesitate to ask your instructors.

Step 2. Research design

During this step, you have to describe all the aspects of your future research. Specifically, you have to explain the following:

- What kind of data you are going to collect? Which variables will be presented in your dataset? How exactly would you collect your data?
- How your hypotheses are stated in statistical terms? What kind of statistical tools (i.e. statistical tests, regression models, etc) would you use to test your hypotheses and answer the research question? Where statistical tests are involved, what is your null hypothesis and what is your alternative?

As a result of first two steps, you have to prepare a document entitled *Research proposal* (one or two pages long) and submit it due **April 25**.

We are going to provide feedback based on your research proposals to guide your research. This means that you don't have to strictly follow your proposed plan: it can be updated according to the feedback, new methods can be introduced that we will discuss after the submission of proposals, etc.

However, you cannot tweak your hypotheses in order to "fit" them to your data, and you cannot discard the data that doesn't support your hypothesis — these are serious violations of research integrity that will invalidate your conclusions.

Step 3. Data collection

It is expected that you do some data collection and/or preprocessing. You are allowed to use data you are collected during some other project (i.e. your BS diploma, etc). You can also use somebody's else data, but in this case your grade will be reduced accordingly, unless you do some additional work that provide additional value to the data (e.g. join several sources of data, do meaningful preprocessing, etc). Anyway, you have to clearly specify the source of the data you use.

The amount of data you use should be large enough to get statistically significant results provided that reasonably large effect actually exists. Usually you need hundreds of observations, more if you use complex multivariate models or your effects are rather small.

In your report, your data should be thoroughly annotated. Specifically, it should be clear how the data was collected, what is your unit of observation and what is the meaning of each variable.

You can use any tools you need during this part, i.e. specific corpus linguistic tools, phonetic data extraction tools, Python, etc. However, if you do data preprocessing, it is preferred if you use R tools.

Step 4. Data analysis

This step usually begins with exploratory data analysis: you provide data visualizations and descriptive statistics that allows to make better understanding of your data. Then you apply the methods you choose, obtain and report quantitative results: p-values, regression coefficients, their significance, etc.

You are expected to use wide range of statistical tools discussed during the course, from simple tests that relate two variables (like t-test or chi-squared) to complex models that take into account several variables (e.g. multivariate linear or logistic regressions, etc.).

You are also encouraged to make robustness checks, if possible, i.e. to show that your conclusions does not change if you slightly change design of your research (e.g. consider several reasonable model specifications, etc.), or explain these changes, if any.

You have to do this step using R.

Step 5. Interpretation

At this step you interpret the results of your quantitative analysis in terms of your research domain. You return to the research question and hypotheses and conclude what can be said about them in the light of your statistical analysis.

Does it support your hypotheses or contradict them? Do you have enough data to make substantial conclusions? How can you explain different results of different statistical tools (if any)? What are possible alternative explanations of your results (if any)?

Step 6. Report

Your report should be Rmarkdown document compilable to HTML and/or PDF that contains full description of your project: the research question and background of your research, your data collection protocol, data description, analysis and interpretation. You have to provide full source code and datasets needed to reproduce your research as well.

To ensure integrity of your research, you have to report difference between the research design you submitted and the actual research conducted.

The final paper should be submitted due **June 20**.

Step 7. Oral presentation

Your exam will consists of open oral presentations of your projects. Each research will get 5 minutes for the talk and 10 minutes for questions. You can prepare slides if you wish, but it is not a requirement, i.e. you can simply use your Rmd report instead of slides. You have to be ready to answer questions about your project as well as theoretical background of all the methods you used.

Assessment

The following aspects of your work will be assessed:

- Research question, hypotheses and background: are they clearly stated?
- Data collection process: is it correct, does it introduce any biases?
- Data description: is it clear what are your observations and variables?
- Exploratory data analysis: does it provide better understanding of your data?
- Choice of statistical tools: are they applicable to your problem and data, do you cover wide range of tools discussed during the course?
- Statistical analysis: is it performed correctly?
- Interpretation of the results in statistical terms: is it correct?
- Interpretation of the results in terms of the research domain: is it correct?
- Presentation and discussion.
- Penalties for late submission of the research proposal or final paper.

Timeline

To summarize, here is a timeline (end of day is 23:59:59 MSK):

- March 23: this document is published.
- April 25: research proposals are submitted.
- June 20: final papers are submitted.
- June 21 — June 30: exam conducted at the date specified by the study office.