Research Project Description

Linguistic data: quantitative analysis and visualization

Possible research tracks and topics In the project, the students are supposed to explore a linguistic phenomenon in written, oral, or sign speech, in any natural language. This can be:

- a corpus-based study:
  - based on a monolingual corpus
  - based on a parallel corpus or any other bi- or multilingual source
- experimental research:
  - psycholinguistic, neurolinguistic experiments
  - experimental phonetics
  - fieldwork
  - sociolinguistic surveys

Just to give you an idea, the project can be focused on the behaviour of some rival forms such as:

- near-synonyms (e.g. *bring* and *take*, the contexts of use of which differ in many ways)
- two or more grammatical forms (e.g. two kinds of comparatives in Russian: *bystree* and *pobystree*)
- rival word formation models
- rival syntactic constructions, word order (e.g. *give him a book* vs. *give a book to your sister*)
- rival pronunciation models (e.g. with or without ellision), etc.

Another example could be the use of dialectal forms in urban and rural speech, or the use of a certain grammatical construction by females and males, or the pronunciation of certain patterns by native and L2 speakers.

For example, in the case of the rival units, the choice between them can be explained by certain contextual variables (e.g. the part of speech of the neighbors, syntactics patterns), features of the rival units themselves (e.g. the gender of the noun, the tense of the verb), genre and register of the text, sociolinguistic parameters (e.g. age, sex, profession, education, place of birth). For inspiration, you can look at some examples of the data sets in the open data repositories Trolling, OSF, PsyArXiv, Kaggle, etc. As another source of motivation, you can look at the description of one of the grammatical categories in Russian RusGram, pick variables mentioned their as associated with/interacting with/explaining the use of the vatiable in focus and make a quantitative case study based on examples from the Russian National Corpus, Aranea, or other available corpora.

It is important to note that we expect you to think about not just one, but several variables that can (potentially) explain the behaviour of the linguistic unit in focus. Please explore at least three explanatory variables of any kind in your case study.

We are ok with the re-use of your previous research but only if it demonstrates some "added" value in terms of statistical analysis, visualization, the increment of the dataset size and variables. If you re-use some previous research, please state it clearly and explain what was done before and what is done as a part of this course project.

Parts of the research

1. A hypothesis formulated in terms of the subject area. For example, *'after verbs in this form, such and such a word is more often used than such and such'*. Or *'people who grew up in the village use dialect*

forms more actively than those who grew up in the city', etc. It would be nice if there is any universal justification for why this hypothesis is plausible.
2. Research design: what data you need to test the hypothesis, what are your formal hypotheses (statistically formulated: null hypothesis, alternative), what is your model (i.e. regression model, etc.), which statistical tools will you use, etc.
3. Description of the data collection method.
4. Collected data, their description.
5. Exploratory data analysis, descriptive statistics and visualization.
6. Results of applying statistical tests and modeling, p-values, regression coefficients, etc.
7. The results of the study, formulated in terms of the subject area.

What usually happens:

During the project, you formulate your initial hypothesis, collect data, annotate data and do the preliminary descriptive, exploratory and inferential statistical analysis. After that, you update your hypothesis, include more or exclude some factors, and collect more data/annotate more parameters in order to improve the empirical basis for the analysis. The amount of the data collected should be enough to support the statistic analysis. (The rule of thumb that the size of your data set is ten times more than the number of variables).

It is strongly prohibited to exclude data that contradict the hypothesis or make any other sort of the hypothesis-biased fraud.

You have to demonstrate robustness (or lack of it) of your conclusions by using different models or their combinations, different specifications of the models (i.e. what variables to include in the analysis), different ways to encode data, etc.

Research paper The students prepare the final project in a written form as an electronic document (Rmarkdown, also knitted as pdf) that include the following parts:

- Research objectives and hypothesis to be tested.
- Description of input data: features and values, descriptive statistics, data visualisation.
- Discussion of the methods of analysis and their applicability.
- Obtained results and their linguistic interpretation. Comparison and discussion of the results produced by different models.
- Optional section: previous research on the topic, comparison of current results with previous studies.
- R code used for the analysis.[1]
- Annotated data (file in repository in the .csv format, the paper contains a link to this file).
- Experimental survey questionaries, if applicable (files in repository in the .csv format, the paper contains a link to this file).

Language under analysis: any natural language

Type of the project: individual or group (max. 2 people) project

Language of the project paper: English

Data annotation

The students can either compile the dataset specifically for this research paper or make use of data collected for their term papers, dissertations, other research actitivities. If you use data collected and annotated by other people please indicate it in your research paper. In this case, you will have to (a) write an additional section on *previous research on the topic, comparison of current results with previous studies* or (b) do additional model testing to explore the effects of data size, missing data treatment, data sampling, custom methods within models, etc.

[1] The use of language R is one of the criteria of evaluation. If you use any statistical tool other than R in your project, please discuss it with the examiner. Python, C, etc. scripts used to collect and pre-process your data can also be provided, but are not subject of evaluation.

Corpus data samples and experimental data sets are usually annotated manually. However, you can do any kind of data preprocessing and exploit ways to automate data annotation, if you wish. The quality of annotation and its interpretability will be assessed.

Project paper and presentation assessment

The following features of your works will be assessed:

- Clear statement of research questions and hypothesis in terms of the subject area
- Research design: relevance of statistical tools used, correctness of statistical assumptions, etc.
- Correctness of data collection method
- Quality of data annotation
- Data description, descriptive statistics and visualization
- Statistical analysis: correctness of statistical tests application, statistical modelling and reporting of results
- Linguistic interpretation of the modeling results, explanation of conflicting results obtained with different models
- Bonus for research quality and extra materials
- Oral presentation. You can prepare additional presentation or just use pictures/tables from your paper
- Penalty for late pre-registration, late final paper submission

Examples of project papers (of different quality, for orientation only).

Please, do not hesitate to contact instructors to discuss research methodologies, data, and any other aspects of your work. We are here to help you!