A Quantitative Analysis of Rival Forms

MA Course: Linguistic data: quantitative analysis and visualization
Final exam project

In the project, the students are supposed to explore the use of the rival forms in the written or oral speech. The rival forms can be:

- two or more synonyms
- two or more grammatical forms
- rival word formation models
- rival syntactic constructions
- rival pronunciation models, etc.

Just to give an idea, the choice of the rival units can be driven by certain contextual factors (e.g. words, 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). We expect students to explore at least **three factors** of any kind in their case study. For inspiration, you can look at some examples of the data annotation at the open repository Trolling (https://opendata.uit.no/dataverse/trolling).

During the project, the students formulate their initial hypothesis, collect data (either corpus-based or experimental), annotate data and do the preliminary descriptive, exploratory and inferential statistical analysis. After that, they can update their hypothesis, include more or exclude some factors, and collect more data/annotate more parameters in order to improve the empirical basis for their analysis. The amount of the data collected should be enough to support the statistic analysis. It is by no means evident that it is strongly prohibited to exclude data that contradict the hypothesis or make any other sort of the hypothesis-biased fraud. At least two multi-factor analysis techniques are two be demonstrated in the paper.

The students are expected to prepare the final project in a written form as an electronic document (R markdown) that include the following parts:

- Research objectives and hypothesis to be tested.
- Description of input data: features and values, descriptive statistics, 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.
- The R code used for the analysis[1].
- Annotated data (as an attachment, preferrably in .csv format).
- Experimental survey questionaries, if applicable (as an attachment).

Language under analysis: any natural language

Language of the project paper: English (Language Theory program), Russian (Computational Linguistics program)

Type of analysis: multi-factor analysis

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

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 etc. scripts used to collect and pre-process your data are not required and will not be evaluated.