Co-funded by the
Erasmus+ Programme
of the European Union

SKILLS

i. Name of the course
Collecting and analyzing corpus data in hypothesis-driven linguistic research: The /k, g, x/ → /t͡s, z, s/ alternation in srWaC

ii. Level of the course
Advanced BA, MA

iii. Workload
3 ECTS

iv. Institution
University of Novi Sad

v. Course instructor(s)
Marko Simonovic

vi. Brief course description

The course provides hands-on experience of corpus data extraction and analysis, targeting a phenomenon which is well known from prescriptive sources, but insufficiently described and virtually unaccounted for in formal approaches. Unlike most cases, where individual students or groups work on a representative sample, in this course the group works as a whole and targets all words attested in the corpus (above a certain frequency threshold).
The course enhances the problem-solving and data-analysis skills, thus preparing the students for a wide range of possible careers. It also provides the students with first-hand scientific research experience.
The course consists of six parts.
During the first, introductory part (10% of the time available), the velar/strident alternation in BCMS is described and the environments in which it occurs are identified. The starting point are the standard grammars. Special attention is devoted to cases of optionality, variation and gaps and the native speakers are encouraged to share their intuitions which diverge from the standard grammars. All students submit a descriptive summary of the velar/strident alternation in BCMS based on available descriptive sources.
The second part (10% of the time available) focuses on general research design and its application to the alternation in focus. Specifically, the various environments in which the alternation applies are translated into independent variables, whereas the application of the alternation is conceptualised as the dependent variable.

In the third part (10% of the time available) the focus is on obtaining data from corpora, specifically, from the Serbian web corpus (srWaC). The students get a quick introduction to CQLs and learn about what kind of data can be obtained.
In the fourth part (20% of the time available), a common project is set up targeting one of the morphological contexts for the alternation (the most probable candidate being the dative/locative singular context). A common document is created and shared with all participants, in which the data collection procedure is described. A Google sheet is created where data get collected. The specifics of the data collection are agreed upon:

- the specific CQLs to be used (+ whether different CQLs are used for triangulation),
- inclusion criteria for lemmas,
    - frequency threshold,
    - word status,
    - unresolvable homonymy etc.
- splitting lemmas (cases of resolvable homonymy),
- merging lemmas (different spellings)
- variables for which the lemmas will be annotated.

The initial division of labour is agreed upon.
Part 5 (40% of the time available) is the central part of the course. In between classes all students do a portion of data collection. The classes serve for discussions of issues and agreeing on changes in the data collection procedures (which get 'registered' in the relevant document). The teacher informs of the descriptive statistics of the data collected up to the point
Part 6 (10% of the time available) is used for a wrap-up of the empirical results.

vii. Research related subject

Conditioning of phonological alternations.

viii. Data the students work with

Data obtained from corpora, descriptive statistics.

ix. Topics

A: Research design

A1: General research design

[Teaching materials]
UPSKILLS Moodle course First steps into scientific research
https://upskillsproject.eu/project/scientific_research/

Movetia/ReLDI courses: PHIL: Movetia101 Introduction to research in linguistics: theory, logic, method https://phil.openedx.uzh.ch/courses/course-v1:PHIL+Movetia101+2046/info (in English) ReLDI-Project: ReLDI101 Introduction to Research Methodology in Linguistics https://phil.openedx.uzh.ch/courses/course-v1:PHIL+ReLDI101+2018/info (in BCMS)
A2: Adapting the general research design to the specific topic of interest
Identifying the predictors of alternation: position in the paradigm, borrowed vs. native, word frequency, phonological environment, animacy etc.
A2.1: Formulation of questions and hypothesis in terms of variables
A2.2: Formulation of predictions of H0 and H1
A2.3: Selection of optimal research techniques, selection and creation of corresponding data sources • Developing and exploiting databases and corpora (e.g. manual data annotation)
A2.4: Identifying the optimal data analysis method
A2.5: Inferring theoretical consequences from the specific data analysis results
A3: Adapting the research design to the available research infrastructures
Familiarising with the type of data extractable from the available corpora
A3.1 Selection of optimal research techniques, selection and creation of corresponding data sources (see also A2.3) • data compilation, data analysis; • understanding, selecting and performing optimal statistical tests and models
B: Infrastructures & techniques
B1: For obtaining, sharing and managing data
Advanced use of srWaC
B1.1: Identifying, collecting, creating and/or using relevant data for research projects • Searching relevant corpora, • Citing linguistic data sets as appropriate
B1.2: Document a research process

B2: For analysing data
B2.1: Concordancers for the analysis of corpora
C: Subject-specific topics
C1: Basic concept of descriptive phonology of BCMS
C2: Basic concept of descriptive morphology of BCMS
C3: Basic concepts of Optimality Theory

x. Learning outcomes

A: Research design
A1: Students will be able to make an overview of the general research design.
[Teaching materials] UPSKILLS Moodle course First steps into scientific research https://upskillsproject.eu/project/scientific_research/ Movetia/ReLDI courses: PHIL: Movetia101 Introduction to research in linguistics: theory, logic, method https://phil.openedx.uzh.ch/courses/course-v1:PHIL+Movetia101+2046/info (in English) ReLDI-Project: ReLDI101 Introduction to Research Methodology in Linguistics https://phil.openedx.uzh.ch/courses/course-v1:PHIL+ReLDI101+2018/info (in BCMS)
A2: Students will be able to create a suitable research design for the specific topic of interest.
Students will be able to identify the predictors of variation and avoidance: position in the paradigm, borrowed vs. native, word frequency, phonological environment, animacy etc.
A2.1: Students will be able to formulate questions and hypothesis in terms of variables.
A2.2: Students will be able to formulate H0 and H1.
A2.3: Students will be able to select optimal research techniques, and create corresponding data sources. • Experimental paradigms (e.g., elicitation, judgements, forced-choice, self-paced reading)

• Developing and exploiting databases and corpora (e.g., manual data annotation).
A2.4: Students will be able to select and implement the optimal data analysis method.
A2.5: Students will be able to infer theoretical consequences from the specific data analysis results.
A3: Students will be able to adapt a research design to the available research infrastructures.
Students will be familiar with the type of data extractable from the available corpora.
A3.1 Students will be able to select of optimal research techniques, select and create corresponding data sources (see also A2.3) • data compilation, data analysis; • understanding, selecting and performing optimal statistical tests and models.
B: Infrastructures & techniques
B1: Students will be able to identify and apply suitable infrastructures & techniques for obtaining, sharing and managing data.
Students will be able to use srWaC in order to extract alternation data.
B1.1: Students will be able to identify, collect, create and/or use relevant data for their research projects • Searching relevant corpora • Citing linguistic data sets as appropriate.
B1.2: Students will be able to document a research process.
B3: Students will be able to identify and apply suitable infrastructures & techniques for analysing data.
B3.1: Students will be able to select and use concordancers for the analysis of corpora.
C: Subject-specific learning outcomes
C1: Students will be able to apply the basic concepts of descriptive phonology to BCMS.
C2: Students will be able to apply the basic concepts of descriptive morphology to BCMS.
C3: Students will be able to apply the basic concepts of Optimality Theory to the alternation in focus.

xi. Overview of evaluation

This course is suitable for pass/fail grading, where all students who regularly contribute to the data collection and in-class discussions pass the course.

Rubric	Weighing
Participation incl. homework (initiative, forward-thinking, problem solving, critical thinking, organisation, time management)	100%