UAB
Universitat Autònoma de Barcelona

Linguistic Data Processing and Analysis

Code: 45505
ECTS Credits: 5

2025/2026

Degree	Type	Year
Advanced Studies in Catalan Language and Literature	OP	1

Contact

Name: Gemma Repiso Puigdelliura

Email: gemma.repiso@uab.cat

Teachers

(External) Josefina Carrera-Sabaté

Teaching groups languages

You can view this information at the end of this document.

Prerequisites

There are no prerequirements

Objectives and Contextualisation

The main objective of this course is to provide students with the methodological tools necessary to carry out rigorous research in the field of linguistics, with a special focus on the study of linguistic variation centered on the phonetic component of language. The course will cover the full cycle of a linguistic research project within the framework of variationist sociolinguistics: from defining the object of study and collecting data, through qualitative and quantitative observation, to statistical analysis and the formalization of results. Students will gain knowledge of the R software, which will be used as the main tool for processing, organizing, and visualizing linguistic data, as well as for applying descriptive and inferential statistical techniques that allow for drawing conclusions from empirical data.

The student should be able to:
• Identify and describe phenomena of linguistic variation.
• Apply quantitative and qualitative research methods in the analysis of linguistic data.
• Use data processing tools, especially the R software, for the manipulation and visualization of linguistic data.
• Formulate hypotheses, perform statistical tests, analyze results, and draw meaningful conclusions from empirical data.
• Present a research proposal and a data analysis report with scientific structure and clear exposition.

Learning Outcomes

1. CA23 (Competence) Apply the knowledge, methods and tools of data collection and processing to create linguistic corpora.
2. CA23 (Competence) Apply the knowledge, methods and tools of data collection and processing to create linguistic corpora.
3. CA24 (Competence) Examine the design of a linguistic research, the processes of selecting informants and the techniques for collecting linguistic data following the principles of research ethics.
4. CA24 (Competence) Examine the design of a linguistic research, the processes of selecting informants and the techniques for collecting linguistic data following the principles of research ethics.
5. KA30 (Knowledge) Identify the phases of collecting, processing, analysing, formalising and presenting linguistic data in a study on the Catalan language.
6. KA31 (Knowledge) Recognise the different specific methodologies to collect linguistic data in a study on the Catalan language.
7. KA31 (Knowledge) Recognise the different specific methodologies to collect linguistic data in a study on the Catalan language.
8. KA32 (Knowledge) Select the appropriate technological tools to manage linguistic corpora in a study on the Catalan language.
9. SA31 (Skill) Conduct different types of statistical and interpretative analyses on linguistic data of the Catalan language in a study.
10. SA32 (Skill) Analyse linguistic data of the Catalan language with the help of specific computer programmes.
11. SA33 (Skill) Make use of primary linguistic sources with digital tools.
12. SA33 (Skill) Make use of primary linguistic sources with digital tools.

Content

The course content is divided into two interrelated areas: the foundations of research in linguistic variation and the tools for statistical processing of linguistic data.

Block 1: Foundations for analyzing language variation

1. Basics for analyzing the internal variation of a language. Production, perception, and speaker subjectivity.

2. Preparation and execution of a research project: research design, conceptual framework, formulation of hypotheses (research questions) and objectives, research relevance, determination of methodological design, timeline.

3. Methodological design: research setting, sample, data collection techniques, and analysis (qualitative and quantitative).

4. Introduction to statistics: definition of variables, types and nature of variables, descriptive and inferential statistics, statistical significance.

Block 2: Processing linguistic data with statistical tools

1. Introduction to R: data preparation-cleaning, transformation, and coding of linguistic variables.

2. Descriptive statistics: data summary and visualization-tables, charts, and measures of central tendency and dispersion.

3. Sampling distributions, sample statistics, and population parameters.

4. Hypothesis testing and parametric tests: design of statistical tests for linguistic variation.

Activities and Methodology

Title	Hours	ECTS	Learning Outcomes
Type: Directed			
Attendance at classes and scheduled activities.	24	0.96	CA23, CA24, KA30, KA31, KA32, SA31, SA32, SA33, CA23
Homework assignments and activities	91	3.64	CA23, SA31, SA32, SA33, CA23

The course combines theoretical sessions with practical and applied activities. Theoretical content will be complemented by in-class exercises, manipulation of real data, and hands-on practice with statistical software. Students will develop a research proposal that integrates the knowledge acquired throughout the course, along with a results report produced using R software.

Annotation: Within the schedule set by the centre or degree programme, 15 minutes of one class will be reserved for students to evaluate their lecturers and their courses or modules through questionnaires.

Assessment

Continous Assessment Activities

Title	Weighting	Hours	ECTS	Learning Outcomes
Practice assignments	20%	2.5	0.1	CA23, KA31, KA32, SA31, SA32, SA33
Research project proposal	40%	3.75	0.15	CA24, KA30, KA31
Statistical analysis exercise	40%	3.75	0.15	KA32, SA31, SA32, SA33

Practical exercises (20%): Students will be assessed on the regular completion and quality of the practical exercises assigned during sessions. These exercises are designed to develop methodological and technical skills related to the analysis of linguistic data and the use of R software. Active participation and engagement in activities will also be considered.

Research project proposal (40%): Students will be required to prepare a formal research project proposal in the field of linguistic variation.

Statistical analysis assignment (40%): Students must submit a final assignment consisting of the statistical analysis of a set of linguistic data using R. This report should include data processing, the application of appropriate statistical techniques, interpretation of results, and the clear and structured presentation of conclusions. Both technical accuracy and argumentative and expository clarity will be evaluated.

Bibliography

Butler, Christopher S. (1985). Statistics in Linguistics. Oxford, Basil Blackwell

Gries, S. T. (2013). *Statistics for linguistics with R: A practical introduction*. Walter de Gruyter.

Labov, W. (2010) Principles of Linguistic Change. Vol 3: Cognitive and Cultural Factors. Malden/Oxford: Wiley-Blackwell.

Moreno-Fernández, F. (2012) Sociolingüística cognitiva. Proposiciones, escolios y debates. Madrid/Frankfurt: Iberoamericana/Vervuert.

Pradilla, M. À. (2008). Sociolingüística de la variació i llengua catalana. Barcelona: Institut d'Estudis Catalans

Sonderegger, M. (2023). *Regression modeling for linguistic data*. MIT Press.

Strelluf, C. (ed.) (2023) The Routledge Handbook of Sociophonetics. Londres: Routledge.

Tagliamonte, S. (2012) Variationist Sociolinguistics. Change, Observation, Interpretation. Malden/Oxford: Wiley-Blackwell

Tatham, M.; Morton, K. (2011). A guide to Speech Production and Perception. Edinburgh University Press: Edimburg.

Verzani, John. (2005). Using R for introductory statistics. Boca Raton: Chapman & Hall.

Vida-Castro, M.; Ávila-Muñoz, A. M. (eds.) (2024) The Continuity of Linguistic Change: Selected Papers in Honour of Juan Andreìs Villena-Ponsoda. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge.

Software

R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.x.x) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Groups and Languages

Please note that this information is provisional until 30 November 2025. You can check it through this link. To consult the language you will need to enter the CODE of the subject.

Name	Group	Language	Semester	Turn
(TEm) Theory (master)	1	Catalan	second semester	afternoon