UNIDADE CURRICULAR LETIVA – OPÇÃO LIVRE
OPTIONAL COURSE UNIT
2022/2023

1. Unidade curricular (UC)/*Curricular Unit*

a) Designação: Análise de dados linguísticos para as Humanidades Digitais

Name: Linguistic Data Analysis for Digital Humanities

b) Número de vagas/*Vacancies*: 20

2. Pequeno texto introdutório que deve refletir, o enquadramento da UC proposta na oferta curricular da NOVA FCSH, bem como, o carácter inovador ou a complementaridade com outras UC's existentes.

O tratamento e a exploração de dados linguísticos digitais, no contexto atual das Humanidades Digitais, é um passo essencial para a prossecução de investigação na área. Para tal, a contribuição da linguística de corpus é incontornável na medida em que, desde sempre, desenvolve os métodos, as ferramentas e o enquadramento teórico que permite extrair informação (linguística e não linguística) de grandes quantidades de dados digitais, atualmente disponíveis em vários repositórios nacionais e internacionais. A UC que aqui se propõe, a desenvolver no contexto do projeto HUGOD - Humanities Going Digital (Erasmus+) do CLUNL, permite dar formação essencial para a investigação atual para estudantes de todas as áreas da NOVA FCSH.

Tendo como base de trabalho dados em formato de texto, objeto de análise que comporta informação relevante sob a forma linguística, os conteúdos programáticos abordados na UC visam permitir uma análise objetiva e informada dos dados, com a ajuda de ferramentas que permitem extrair informação de grandes quantidades de dados digitais. Para tal, é necessário perceber como constituir conjuntos de dados relevantes e que métodos de análise permitem chegar a resultados sólidos, recorrendo tanto a critérios linguísticos, estatísticos e outros. A introdução teórica à área da linguística de corpus permite esta perceção, bem como o conhecimento de ferramentas informáticas disponíveis. A ligação entre dados linguísticos e informação requer ainda a sensibilização para a análise linguística e treino no mapeamento entre formas e estruturas linguísticas e a informação específica a extrair, também trabalhados nos conteúdos da UC, bem como a consolidação das competências teóricas adquiridas através de trabalho prático.

The treatment and the exploration of digital linguistic data, in the current context of Digital Humanities, is an essential step to pursue research in the field. The contribution of Corpus Linguistics to that end is unavoidable since, from its conception, Corpus Linguistics develops the methods, tools and theoretical framework that allows for extracting (linguistic and non-linguistic)

information form large quantities of digital txt data, currently available in several national and international repositories. The curricular unit proposed here, developed in the context of the HUGOD - Humanities Going Digital project (Erasmus+) ongoing at CLUNL, allows for providing training essential to current research to students from all the NOVA FCSH domains.

Having as basis textual data, an object of analysis that includes relevant information in linguistic form, the contents covered in the course aim at allowing for an objective and informed analysis of texts, with the help of tools that are used to extract information from large amounts of digital data. To achieve this, students need to know how to select or compile relevant data sets and which methods of analysis will allow them to achieve solid results, considering linguistic, statistical, and other criteria. The theoretical introduction to the field of Corpus Linguistics provides them this perception, as well as knowledge on current and available computational tools and methods. The relation between linguistic data and information requires promoting the sensibility to linguistic data and phenomena and the training in the mapping between linguistic forms and structures and the specific information they intent to extract from data, also covered by the course. The consolidation of the theoretical skills acquired through real hands-on work is also provided.

3. Código da unidade curricular/*Curricular unit code*: *[Não Preencher]*

4. Faculdade/*Faculty*: Faculdade de Ciências Sociais e Humanas

5. Unidade de Investigação/*Research Unit*: Centro de Linguística da Universidade NOVA de Lisboa (https://clunl.fcsh.unl.pt/)

6. Curso/*Course*: Opção livre aberta a todos os cursos de mestrado

7. Nível do curso/*Course Level*: Mestrado

8. Carácter da unidade curricular: Opcional/*Optional*

9. Tipo da unidade curricular/*Type of curricular unit*: Unidade Curricular Letiva

10. Percentagem de aulas práticas/*Percentage of practical classes*: 50%

11. Ano do plano de estudos/*Syllabus year*: 1.º ano

12. Semestre/*Semester*: 1.º semestre/1st semester

13. Número de créditos/*Number of credits* (1 crédito = 28h): 10 ECTS (mestrado)

14. Docente ou Investigador responsável/*Teacher or principal researcher*: Raquel Amaro (responsável); Chiara Barbero e Sílvia Barbosa (investigadoras do CLUNL)

15. a) Número de horas por sessão/*Number of hours per session*: 3 horas (mestrado)

b) Número de sessões por semestre/*Number of sessions per semester*: 16 sessões (mestrado)

c) Periodicidade/*periodicity*: Semanal

d) Período de funcionamento/*Class period*: 19 de setembro de 2022 - 13 de janeiro de 2023

16. Objetivos da unidade curricular/*Learning objectives*:

. Conhecer, compreender e avaliar os métodos e as ferramentas de análise e de extração de informação de grandes conjuntos de dados linguísticos.

. Saber como organizar e utilizar dados linguísticos para extração de informação direcionada e útil para questões de investigação específicas da área das Humanidades Digitais.

. Conhecer métodos de análise e deteção de pistas e traços linguísticos e determinar qual a sua relevância para a extração de informação específica ou para tarefas de mineração de texto para fins não-linguísticos.

. Desenvolver competências para construir e usar corpora textuais de modo analítico e crítico de acordo com metodologias testadas e através de ferramentas de tratamento e análise de corpus.

. Desenvolver competências e estratégias de deteção e utilização de pistas e traços linguísticos para fins de investigação em Artes e Humanidades (Digitais).

. Get acquainted with, to understand and to evaluate the methods and tools for analyzing and extracting information from large sets of linguistic data.

. Know how to organize and use large sets of linguistic data to extract useful information directed and relevant to research issues specific of the Digital Humanities field.

. Know methods of analysis and detection of linguistic cues and features and to determine their relevance to perform specific information extraction and text mining tasks for non-linguistic purposes.

. Develop skills to build and use textual corpora in and analytical and informed way, according to tested methodologies and using available tools for corpus treatment and analysis.

. Develop skills and strategies for detecting and using linguistic cues and features for research purposes in the field of (Digital) Arts and Humanities.

17. Competências gerais do grau/*General skills of the degree*: a); b); c); d) ;e); f)

18. Competências específicas do curso/*Specific Course skills*: Não aplicável./Not applicable.

19. Requisitos de frequência/*Attendance requirements*: Não aplicável / Not applicable.

20. Conteúdo da unidade curricular/*Syllabus* (máx. 200 palavras):

1. Linguística de corpus

1.1. Introdução e enquadramento teórico

1.2 Constituição de corpus: critérios, parâmetros e representatividade

1.3 Ferramentas e procedimentos para tratamento de corpus: panorama geral

2. Dos dados linguísticos à estação de informação específica

2.1 Unidades, traços e pistas linguísticas

2.2 Análise de textos: nível macro vs. nível micro; análise sintomática vs. análise paradigmática

2.3 Estatística lexical, concordâncias e collocations

3. Aplicação de estratégias da linguística de corpus e de text mining

3.1 Objetivos de investigação, seleção de dados e compilação do corpus

3.2 Determinação de pistas e traços linguísticos relevantes

3.3 Extração e análises de resultados

1. Corpus Linguistics

1.1. Introduction and theoretical framework

1.2 Corpus constitution: criteria, parameters and representativeness

1.3 Corpus tools and procedures: overview

2. From linguistic data to specific information extraction

2.1 Linguistic units, features and cues

2.2 Texts analysis: macro vs. micro level; linear/phrasal vs. paradigmatic analysis

2.3 Lexical statistics, concordances and collocations

3. Applying Corpus Linguistics and text mining strategies

3.1 Research question, data selection and corpus compilation

3.2 Determining relevant linguistic features and cues

3.3 Results extraction and analysis

21. Bibliografia recomendada/*Recommended reading*: (máx. 5 títulos. Por ordem decrescente de data de edição.)

1. Hinrichs, E. M. Hinrichs, S. Kübler & T. Trippel (eds.) (2019). Language Resources and Evaluation: Language Technologies for Digital Humanities 53(4). https://doi.org/10.1007/s10579-019-09482-4

2. Odebrecht, C., Belz, M., Zeldes, A., Lüdeling, A. & Krause, T. (2017). RIDGES Herbology: Designing a Diachronic Multi-Layer Corpus. In: Language Resources and Evaluation 51.3, pp. 695–

725.

3. Beloso, B. S. (2015). Designing, Describing and Compiling a Corpus for English Architecture. In Procedia - Social and Behavioral Sciences 198. Elsevier. 459-464.

4. Ebensgaard Jensen, K. (2014). Linguistics and the digital humanities: (Computational) corpus linguistics. MedieKultur: Journal of Media and Communication Research, 30, pp. 117-136

5. McEnery, T. & A. Hardie (2012). Corpus Linguistics: Method, theory and practice. Cambridge University Press.

22. Métodos de ensino/*Teaching Methods*:

O curso conjugará exposição teórica com trabalho prático, favorecendo a descoberta ascendente de necessidades teóricas e metodológicas e de resultados relevantes orientadora para as Artes e as Humanidades.

Tal significa atividades controladas que permitam aos estudantes:

- trabalhar os dados, usando as ferramentas e os métodos necessários,
- obter os resultados esperados, compreendendo os pressupostos e as explicações teóricas emanadas da linguística de corpus,
- chegar a conclusões e propor novo conhecimento nas áreas das Artes e Humanidades, com base na análise de grandes quantidades de dados linguísticos digitais,
- analisar e confrontar as novas propostas com o estado da arte, conhecendo o potencial e as limitações da utilização de métodos e ferramentas computacionais para a exploração de dados linguísticos com objetivo de investigação das Artes e Humanidades.

Os métodos e as atividades de ensino incluirão exposição teórica (através de aulas e de leituras autónomas), atividades práticas orientadas (auxiliadas por guiões, vídeos de demonstração, aulas práticas), discussão e cooperação entre pares, avaliação e monitorização dos processos de trabalho e aprendizagem (através de questionários), redação e apresentação de artigos/ensaios.

The course will be delivered conjugating theoretical exposure and hands-on work, favoring the bottom-up discovery of theoretical and methodological needs and of relevant Arts and Humanities-oriented results.

This means controlled activities that will allow students to:

- work the data, using all the necessary tools and methods
- achieve the expected results, understanding the theoretical assumptions and explanations

provided by Corpus Linguistics

- reach conclusions and propose new knowledge in Arts and Humanities, based on the analysis of large amounts of linguistic digital data
- analyze and confront the new proposals with the state of the art, being aware of the potential and the limitations of using computational methods and tools for exploring linguistic data for Arts and Humanities research goals.

The teaching methods and activities will include theoretical exposure (through classes and autonomous readings), guided hands-on activities (aided by scripts, demonstration videos, hands-on classes), peer discussion and cooperation, workflow evaluation and monitoring (through questionnaires), paper/essay writing and presentation.

23. Métodos de avaliação/*Assessment methods*:

Avaliação contínua, incluindo os seguintes elementos de avaliação: participação ativa nas atividades do seminário (30%) e trabalho de projeto, a desenvolver ao longo do semestre (70%).

Continuous assessment, including the following assessment elements: active participation in seminar activities (30%) and a project work, to be developed throughout the semester (70%).

24. Língua de ensino/*Teaching language*: Inglês/English