**Encuadre de la asignatura**

|  |  |  |
| --- | --- | --- |
| **Datos de identificación** | | |
| Unidad académica: Facultad de Idiomas | | |
| Programa: Doctorado en Ciencias del Lenguaje | | Plan de estudios: |
| Nombre de la unidad de aprendizaje: Procesamiento de datos lingüísticos | | |
| **Perfil de egreso del programa** | | |
| El programa de Doctorado en Ciencias del Lenguaje (DCL) tiene como propósito principal el formar investigadores de alto nivel académico en el campo de las lenguas para producir investigación original de gran consistencia teórica y rigor metodológico de acuerdo con lo siguiente:   * Diseñar, realizar y evaluar proyectos de investigación, en especial en el ámbito de su campo disciplinario y en lo general en perspectiva de trabajo transdisciplinario. * Coordinar, dirigir y asesorar a grupos de investigación en instituciones de enseñanza y de investigación en lenguas del país. * Dominar los fundamentos teóricos y metodológicos del campo de conocimiento de su formación para resolver problemas de manera colaborativa con especialistas de diversas disciplinas que realicen proyectos de investigación originales. * Generar conocimientos sólidos y originales en el campo en que se inserta su línea de investigación. * Desarrollar investigación con un alto compromiso social y humano en beneficio de la solución de problemas de sectores menos favorecidos. * Procesar, analizar e interpretar datos lingüísticos y de ambientes de enseñanza y aprendizaje para interpretar y derivar a elaboraciones de corpus lingüísticos, análisis de fenómenos lingüísticos, diseños curriculares, enfoques y metodologías didácticas en lenguas.   Analizar y utilizar recursos tecnológicos con fines de aplicación lingüística y de enseñanza y aprendizaje para el desarrollo tecnológico y la innovación en el campo de las lenguas. | | |
| **Definiciones generales de la unidad de aprendizaje** | | |
| **Propósito general de esta unidad de aprendizaje:** | Esta Unidad de Aprendizaje tiene como objetivo proporcionar los fundamentos teóricos y metodológicos, propios de la lingüística de corpus y del procesamiento del lenguaje natural, para el procesamientos de datos lingüísticos. Se hará particular énfasis en los aspectos metodológicos y en las situaciones prácticas, principalmente, las relacionadas con aspectos descriptivos del lenguaje, de manera que el alumno sea capaz de i) construir corpus de forma automática o semiautomática, ii) procesar los datos del corpus para extraer información, iii) utilizar software especializado para explotar los corpus y iv) generar explicaciones plausibles con base en la aplicación de herramientas, tales como concordancias, colocaciones, coligaciones, así como de representaciones estadísticas, tales como frecuencias, X2 , información mutua, logaritmo de verosimilitud, entre otras. | |
| **Objetivo de la unidad de aprendizaje:** | Utilizar los principales métodos y herramientas de la Lingüística de Corpus (LC) y algunos del Procesamiento del Lenguaje Natural (PLN) para la obtención, procesamiento y estructuración de datos, mediante el uso de software especializado y la aplicación de métricas estadísticas manteniendo un perfil ético, analítico y responsable en el proceso. | |
| **Evidencia de aprendizaje (desempeño o producto a evaluar) de la unidad de aprendizaje:** | Entrega de una carpeta de trabajo en la que se deben registrar las siguientes evidencias:  1. Compilación de un corpus con datos textuales representativos del fenómeno que le interesa investigar.  2. Reporte del trabajo con herramientas de la LC para el análisis de datos: KWIC, lexicología, terminología, así como de aplicación de estándares para eliminar la información irrelevante.  3. Reporte del trabajo con herramientas de PLN para el etiquetado de datos considerando diferentes niveles (fonético, morfológico, sintáctico, semántico o referencial), así como de la aplicación de métricas estadísticas para la argumentación de resultados.  4. Presentación de resultados. | |

|  |  |
| --- | --- |
| **Temario** | |
| **I. Nombre de la unidad: Fundamento y diseño de corpus** | **Horas:** |
| **Objetivo de la unidad:**  Comprender los conceptos básicos de corpus lingüísticos y desarrollar habilidades para diseñarlos y estructurarlos. | |
| **Tema y subtemas:**   1. Introducción a los corpus lingüísticos: tipos y aplicaciones 2. Diseño de corpus: principios, recopilación y selección de datos 3. Ética y manejo de datos sensibles 4. Herramientas para la recopilación y organización inicial de datos: ELAN y herramientas online | |
| **Prácticas de taller:**  1. Con base en los objetivos del proyecto doctoral del alumno, definir qué tipo de dato lingüístico es idóneo o necesario para el desarrollo del proyecto.  2. En práctica de campo, recopilar un conjunto pequeño de datos a partir de entrevistas o mediante el registro personal de información lingüística con el fin de evaluar qué tipo de conocimiento se podría generar con esos datos.  3. En práctica en el laboratorio de cómputo, analizar diferentes tipos de datos a nivel diacrónico, sincrónico, monolingüe y multilingüe para observar su sistematización en corpus electrónicos. | **Horas:** |

|  |  |
| --- | --- |
| **II. Nombre de la unidad: Herramientas de la LC para el procesamiento de datos** | **Horas:** |
| **Objetivo de la unidad:** Aplicar un conjunto de herramientas de la LC para el procesamiento de corpus y familiarizarse con los procesos de codificación y anotación. | |
| **Tema y subtemas:**   1. Herramientas especializadas    1. Praat para análisis fonético    2. ELAN para anotación multimodal    3. AntConc para análisis textual 2. KWIC, colocaciones y coligaciones 3. Codificación y anotación de corpus lingüísticos (manual y automática) 4. Introducción a lenguajes de programación para análisis (Python con NLTK, spaCy o pandas) | |
| **Prácticas de taller:**  1. Explorar en el laboratorio de cómputo las herramientas descritas y seleccionar una para construir un corpus de forma automática.  2. Conocer algunas herramientas de software libre y propietario para identificar y extraer automáticamente KWIC, terminología, lexicología y fraseología, aplicando utilidades y estándares para la eliminación de información irrelevante. | **Horas:** |

|  |  |
| --- | --- |
| **III. Nombre de la unidad: Herramientas del PLN para el análisis de corpus** | **Horas:** |
| **Objetivo de la unidad:** Aplicar métodos y herramientas del PLN utilizando software especializado de análisis. | |
| **Tema y subtemas:**   1. Herramientas para análisis multifactorial de datos    1. Introducción a Iramuteq 2. Codificación de corpus y especificidades 3. Clasificación jerarquizada y similitud semántica 4. Graficación: análisis factorial de correspondencias | |
| **Prácticas de taller:**  1. Instalar Iramuteq y etiquetar los datos del corpus.  2. Establecer la codificación del corpus de acuerdo con las necesidades del proyecto.  3. Extraer las especificidades del corpus.  4. Realizar diferentes experimentos de clasificación y similitud semántica  5. Representar las caracterísitcas más sobresalientes del corpus mediante AFC. | **Horas:** |

|  |  |
| --- | --- |
| **IV. Nombre de la unidad: Métodos estadísticos para la interpretación de datos lingüísticos** | **Horas:** |
| **Objetivo de la unidad:** Aplicar métricas de análisis estadístico para explicar el comportamiento de los datos y argumentar resultados con base en elementos estadísticos. | |
| **Tema y subtemas:**   1. Métodos cuantitativos para el análisis de datos    1. Fundamentos de Estadística descriptiva    2. Fundamentos de Probabilidad y Estadística inferencial 2. Herramientas para análisis cuantitativo    1. Excel avanzado para análisis estadístico    2. Aplicaciones de macros en Excel 3. Interpretación de resultados estadísticos y lingüísticos    1. Contraste con hipótesis y objetivos iniciales    2. Tabulación, graficación y presentación contextualizada | |
| **Prácticas de taller:**  1. Hacer cálculos estadísticos sobre los datos del corpus para observar el comportamiento de los mismos.  2. Realizar la interpretación del comportamiento de los datos en función de algunas métricas estadísticas.  3. Complementar el análisis con base en fundamentos lingüísticos. | **Horas:** |

|  |
| --- |
| **Estrategias de aprendizaje y de enseñanza utilizadas**:   * 1. Aquí habría que poner lo de la parte intensiva del curso, ¿no? |
| **Criterios de evaluación:**   1. Entrega de la carpeta de trabajo, en la que cada evidencia tiene el siguiente porcentaje: 2. Compilación de corpus: 25% 3. Reporte de herramientas de LC para la explotación de corpus: 25% 4. Reporte de herramientas del PLN para el análisis de datos y aplicación de métricas estadísticas: 25% 5. Presentación de resultados: 25%   **Criterios de acreditación:**   1. Entregar las evidencias en los tiempos establecidos, considerando los valores éticos y profesionales de la comunidad universitaria, así como la normatividad antiplagio del posgrado y de la UABC. 2. Realizar las actividades en las fechas programadas. 3. Tener como promedio general una calificación igual o mayor a 7. |
| **Bibliografía:**   * Callies, Marcus, and Sandra Götz. (2015). *Learner Corpora in Language Testing and Assessment*. John Benjamins. * Corpas Pastor, Gloria, and Jean-Pierre Colson (2020). Computational Phraseology. John Benjamins. Doval, Irene, and M. Teresa Sánchez Nieto. (2019). Parallel Corpora for Contrastive and Translation Studies. John Benjamins. * Clark, Alexander, Chris Fox and Shalom Lappin. (2010). *The Handbook of Computational Linguistics and Natural Language Processing*. Wiley-Blackwell. “Clásica”. * Corpas Pastor, Gloria, and Jean-Pierre Colson (2020). *Computational Phraseology*. John Benjamins. * Doval, Irene, and M. Teresa Sánchez Nieto. (2019). *Parallel Corpora for Contrastive and Translation Studies*. John Benjamins. * Gómez-Pérez, J., R. Denaux, and A. Garcia-Silva. (2020). *A Practical Guide to Hybrid Natural Language Processing*. Springer International Publishing. * Jurafsky, Dan and James Martin. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (Third edition, August 2020)*. Prentice Hall. Disponible en <https://web.stanford.edu/~jurafsky/slp3/>. “Clásica”. * Manning, Chris, (2014). *Statistical natural language processing and corpus-based computational linguistics: An annotated list of resources*. Standford Universty. Disponible en <https://nlp.stanford.edu/links/statnlp.html>. “Clásica”. * McEnery T, Brezina V. (2022). *Fundamental Principles of Corpus Linguistics*. Cambridge University Press. * Mitkov, Ruslan (2022). *The Oxford Handbook of Computational Linguistics (Second edition, June 2022).* Oxford University Press*.* |
| **Fecha de elaboración / actualización:**  **Actualización: 13 de enero de 2025** |