FACULTAD DE

**LETRAS Y CIENCIAS**

**HUMANAS**

**ESPECIALIDAD: LINGÜÍSTICA Y LITERATURA**

**Nombre del curso: Lingüística computacional**

**Código del curso: 1LIN-11**

**Tipo de curso: Electivo**

**Año: 2021**

**Semestre: Segundo**

**Número de créditos: Cuatro (04)**

**Requisito: No tiene**

**Profesor del curso: Roberto Zariquiey Biondi**

**Javier Vera Zúñiga**

1. **Sumilla**

El curso ofrece una introducción a la lingüística computacional desde dos perspectivas complementarias: teórica y práctica. La primera discute la necesidad de su multidisciplinariedad para dar cuenta de fenómenos como el procesamiento del lenguaje o las tecnologías del lenguaje humano. La segunda ofrece un panorama de las tecnologías que actualmente subyacen a los últimos avances en el tratamiento del lenguaje natural en la industria de la informática.

1. **Enfoque temático**

Exploraremos los avances más recientes en lingüística computacional, prestando atención a temas como procesamiento y modelaje de lenguas naturales, a partir de principios estadísticos y/o reglas generales. El foco se encuentra en el uso de estas técnicas para el estudio de la diversidad de las lenguas amerindias desde una perspectiva tipológica. En esa línea, nuestro curso se enfoca en los métodos computacionales elementales con los que se pueda promover la investigación a gran escala, en tipología y áreas lingüísticas, con foco en lenguas sudamericanas. El curso se construye desde dos perspectivas: un enfoque teórico, que expone las principales ideas y preguntas sobre aproximaciones computacionales al lenguaje humano; y un enfoque práctico, que permite a los estudiantes adentrarse en el desarrollo de tecnologías simples para el análisis automático de datos lingüísticos.

1. **Sistema de evaluación**

La evaluación está conformada por:

* Examen Parcial (30%)
* Examen Final (30%)
* Tareas de implementación computacional (40%)

1. **Contenidos**

* Introducción: el rol de los métodos computacionales en las nuevas ciencias del lenguaje: tipología, arealidad, descripción gramatical y NLP
  + Nuevas preguntas: ¿Qué problemas son abordables con un computador? ¿Qué nuevas perspectivas se abren?
  + Disciplinas relacionadas: Lingüística Computacional, Procesamiento del Lenguaje Natural y Análisis de Datos.
  + Lingüística computacional y lenguas minoritarias: computarización de lenguas
  + El proyecto Chana

.

* Bases de datos lingüísticas: Glottolog, SAILS, WALS, Glotobank, corpus paralelos, datos sobre obsolescencia.
  + Tipos de datos lingüísticos: ¿Qué es un dato lingüístico? Perspectivas éticas de datos y lenguas.
  + Exploraciones en bases de datos.
  + Datos y lenguas de Sudamérica.
* Nociones de programación con Python
  + ¿Dónde programar?: Jupyter Notebooks y plataformas para Python.
  + Elementos básicos: variables, listas, diccionarios, funciones.
  + Estructuras de control. Nociones básicas de lógica para computación.
  + Manejo computacional de datos lingüísticos.
  + Compartir proyectos a través de GITHUB.
* Visualización de información lingüística.
  + Construcción de mapas.
  + Gráficos y presentación de resultados.
  + Uso de bases de datos para la extracción de información sobre lenguas peruanas y Sudamericanas.
* Manipulación de bases de datos lingüísticas.
  + Herramientas básicas de análisis: estadística descriptiva.
  + Distancias entre lenguas.
  + Problemas sobre frecuencia de palabras: types y tokens.

1. **Competencias y resultados de aprendizaje**

Los objetivos del curso son los siguientes:

* Conocer el carácter multidisciplinario de aproximaciones computacionales al lenguaje humano.
* Conocer las principales líneas de trabajo con respecto a datos lingüísticos.
* Adentrarse en la implementación computacional de problemas simples que involucren el estudio computacional del lenguaje humano.
* Conocer las principales fuentes de datos sobre lenguas, en especial de Perú y de Sudamérica.
* Adquirir competencias básicas sobre análisis de datos y manejo de herramientas computacionales.

1. **Método de trabajo**

El curso se realizará mediante co-dictado en modalidad no presencial, utilizando esencialmente sesiones de zoom. El trabajo se basa en clases expositivas con alto contenido práctico, a cargo de los profesores, el desarrollo de tareas guiadas de implementación computacional y la elaboración de un proyecto computacional final por parte de los estudiantes. Los estudiantes deberán contar con una computadora personal en donde deberán instalar los programas y los entornos de funcionamiento indicados por los profesores.

1. **Bibliografía general**

Bickel, Balthasar. 2011. [Statistical modeling of language universals.](http://www.zora.uzh.ch/57330) Linguistic Typology, 15(2):401 - 414.

Bickel, Balthasar (2015). [Distributional typology: statistical inquiries into the dynamics of linguistic diversity.](http://www.zora.uzh.ch/109110) En Heine, Bernd; Narrog, Heiko (editores), The Oxford handbook of linguistic analysis, 2nd edition. Oxford: Oxford University Press, 901 - 923.

Blasi, Damián E., Søren Wichmannd, Harald Hammarström, Peter F. Stadler, y Morten H. Christiansenh. 2014. Sound–meaning association biases evidenced across thousands of languages. PNAS. www.pnas.org/cgi/doi/10.1073/pnas.1605782113

Blasi, Damián E., Susanne Maria Michaelis y Martin Haspelmath. 2017. Grammars are robustly transmitted even during the emergence of creole languages. Nature Human Behaviour, letters.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard y Quentin D. Atkinson. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. Science 337, 957.

Dingemanse, Mark, Damián E. Blasi, Gary Lupyan, Morten H. Christiansen y Padraic Monaghan. 2015. Arbitrariness, Iconicity, and Systematicity in Language. Trends in Cognitive Sciences 19 (10).

Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, Accessed on 2021-06-09.)

Hammond, Michael. Python for Linguists. Cambridge University Press, 2020.

Hammarström, Harald y Tom Güldemann. 2014. Quantifying Geographical Determinants of Large-Scale Distributions of Linguistic Features. En Søren Wichmann y Jeff Good (editores) Quantifying Language Dynamics: On the Cutting Edge of Areal and Phylogenetic Linguistics. Leiden: Brill.

Michael, Lev, Will Chang and Tammy Stark. 2014. Exploring Phonological Areality in the Circum-Andean Region Using a Naive Bayes Classifier. En Søren Wichmann y Jeff Good (editores) Quantifying Language Dynamics: On the Cutting Edge of Areal and Phylogenetic Linguistics. Leiden: Brill.

Michael, Lev, Natalia Chousou-Polydouri, Keith Bartolomei, Erin Donnelly, Vivian Wauters, Sérgio Meira, Zachary O’Hagan. 2015. A Bayesian Phylogenetic Classification of Tupí-Guaraní. LIAMES 15(2), 193-221.

Muysken, Pieter, Harald Hammarström, Olga Krasnoukhova, Neele Müller, Joshua Birchall, Simon van de Kerke, Loretta O'Connor, Swintha Danielsen, Rik van Gijn & George Saad. 2016. South American Indigenous Language Structures (SAILS) Online. Jena: Max Planck Institute for the Science of Human History. (Available at https://sails.clld.org)

Proyecto Chana. 2018. CLD3: Computational Language Documentation, Description and Development. Lima: PUCP (documento interno).

Seifart, Frank. 2014. Does structural-typological similarity affect borrowability? A quantitative study on affix borrowing. En Søren Wichmann y Jeff Good (editores) Quantifying Language Dynamics: On the Cutting Edge of Areal and Phylogenetic Linguistics. Leiden: Brill.

Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python. O’Reilly Media, Inc., 1st edition, 2009.

Xia, Fei and William Lewis. 2008. Repurposing theoretical linguistic data for tool development and search. Proceedings of IJCNLP-2008. Hyderabad, India.

1. **Cronograma**

| **Semestre 2021-2** | | | **Tema de Clase** | **Evaluaciones, salidas de campo y otras fechas** | **Actividades en Paideia** |
| --- | --- | --- | --- | --- | --- |
| **MES** | **SEM** | **FECHAS** |
| **AGOSTO** | **1°** | **23-28** | Introducción a la Lingüística Computacional |  | Sesiones sincrónicas por Zoom  Lecturas  Foro |
| **AGOSTO-SETIEMBRE** | **2°** | **30-04** | Datos lingüísticos: tipos y consideraciones éticas | **Feriado: lunes 30** | Sesiones sincrónicas por Zoom  Lecturas  Foro |
| **3°** | **06-11** | Nociones básicas de Python |  | Sesiones sincrónicas por Zoom  Lecturas  Programación |
| **4º** | **13-18** | Github |  | Sesiones sincrónicas por Zoom  Lecturas  Programación |
| **5º** | **20-25** | Python y manejo de datos lingüísticos |  | Sesiones sincrónicas por Zoom  Lecturas  Programación |
| **SETIEMBRE-OCTUBRE** | **6º** | **27-02** | Bases de datos lingüísticas: lenguas Peruanas y de Sudamérica |  | Sesiones sincrónicas por Zoom  Lecturas  Programación |
| **7º** | **04-09** | Python y manejo de datos lingüísticos | **Feriado: viernes 8** | Sesiones sincrónicas por Zoom  Lecturas  Programación |
| **8º** | **11-16** | Gráficos y presentación de resultados |  | Sesiones sincrónicas por Zoom  Lecturas  Programación |
| **9º** | **18-23** | **suspensión de clases y exámenes de medio ciclo** | | |
| **OCTUBRE-NOVIEMBRE** | **10º** | **25-30** | Panorama sobre lenguas Peruanas y Sudamericanas |  | Sesiones sincrónicas por Zoom  Lecturas  Programación |
| **11º** | **01-06** | Extracción de información sobre lenguas Peruanas y Sudamericanas | **Feriado: lunes 1** | Sesiones sincrónicas por Zoom  Lecturas  Programación |
| **12º** | **08-13** | Estadística descriptiva sobre lenguas Peruanas y Sudamericanas |  | Sesiones sincrónicas por Zoom  Lecturas  Programación |
| **13º** | **15-20** | Construcción de mapas lingüísticos | **Inicio de Encuesta virtual de opinión sobre docentes: viernes 19** | Sesiones sincrónicas por Zoom  Lecturas  Programación |
| **14º** | **22-27** | Problemas sobre frecuencias de palabras |  | Sesiones sincrónicas por Zoom  Lecturas  Programación |
| **NOVIEMBRE-DICIEMBRE** | **15°** | **29-04** | Información tipológica y distancias entre lenguas | **Fin de Encuesta virtual de opinión sobre docentes: lunes 29** | Sesiones sincrónicas por Zoom  Lecturas  Programación |
| **16º** | **06-11** | **clases y exámenes de fin de ciclo** | | **Feriado: miércoles 8** |
| **17º** | **13-15** | **exámenes de fin de ciclo (solo lunes, martes y miércoles)** | | |
| **16-17** | **Exámenes de rezagados (solo jueves y viernes)** | | |
|  |  |  | **FECHA FINAL DE ENTREGA DE NOTAS: miércoles 22 de diciembre** | | |

**\* IMPORTANTE:**El calendario del curso podrá modificarse si la marcha de la asignatura lo requiere. En el caso en que estos cambios impliquen cambios de fechas de actividades académicas (como evaluaciones y otros), el docente dará oportuno aviso a los estudiantes, a la Dirección de Estudios y a la Secretaría Académica de la Facultad. En ningún caso se modificará la sumilla del curso.

**\***Conforme a los lineamientos establecidos por el Ministerio de Educación y la Superintendencia de Nacional de Educación Superior Universitaria (SUNEDU) dictados en el marco de la emergencia sanitaria para prevenir y controlar el COVID-19, la universidad ha decidido iniciar las clases bajo la modalidad virtual hasta que por disposición del gobierno y las autoridades competentes se pueda retornar a las clases de modo presencial. Esto involucra que los docentes puedan hacer los ajustes que resulten pertinentes al sílabo atendiendo al contexto en el que se imparten las clases.