¿Qué es la Lingüística Computacional (LC)?

La lingüística computacional es el estudio científico del lenguaje humano o natural desde una perspectiva computacional. Es un campo interdisciplinario en desarrollo que abarca la lingüística teórica, el procesamiento de lenguas naturales, las ciencias de computación, la inteligencia artificial, la psicología, la filosofía, las matemáticas y la estadística, entre otras. Los lingüistas computacionales se interesan en proveer modelos computacionales para varios tipos de fenómenos lingüísticos. Estos modelos pueden ser de **conocimiento** (basados en conocimiento del mundo y competencia lingüística) o **estocásticos** (basados en probabilidad y estadística a partir de datos). La investigación en lingüística computacional está motivada en algunos casos desde una perspectiva científica, en la que se trata de ofrecer una explicación computacional para un fenómeno lingüístico o psicolingüístico en particular; en otros casos la motivación puede ser más bien tecnológica, en la que se quiere proveer un componente operacional para un sistema de habla o lenguaje natural. De hecho, el trabajo de los lingüistas computacionales se incorpora en muchos sistemas operacionales, incluyendo sistemas de reconocimiento de habla, sintetizadores de texto a habla, sistemas de respuesta de voz automática, motores de búsqueda en la red, editores de texto y materiales de instrucción de lenguas, entre otros.

Para más información, visite el Departamento de Estudios Hispánicos (Chardón 503, L-V 7:30 am – 4:30 pm) 787.265.3843 o 787.832.4040 ext. 3843

¿Qué aplicaciones tiene la Lingüística Computacional?

Los trabajos en lingüística computacional tienen aplicaciones tecnológicas cada vez más necesarias y cotizadas en la industria:

- ***interfaces humano-máquina***, (agentes conversacionales), en las que se utiliza una lengua natural en vez de una artificial o un menú restringido de opciones
- ***reconocimiento y síntesis de habla y texto*** (sistemas de respuesta, sintetizadores de voz y transcriptores), que requieren conocimiento sintáctico para procesar aspectos prosódicos (entonación)
- ***traducción automática*** a otras lenguas a partir de producciones textuales u orales
- ***motores de búsqueda y recuperación de información***, que requieren entender las condiciones de búsqueda para reconocer qué documentos son relevantes o no
- ***extracción de información***, que necesitan reconocer la información relevante en una base de datos para trasladarla a formatos predeterminados (como tablas o gráficas)
- ***entrañamiento textual***, que requieren entendimiento de lenguaje natural para reconocer inferencias y verificar hipótesis a partir de textos diversos.
- ***correctores gramaticales y de estilo***, que necesitan conocimiento sintáctico para detectar fallos de concordancia y oraciones "incompletas" o "incorrectas"
- ***correctores ortográficos***, que deben poseer al menos conocimiento de análisis morfológico y estructura silábica
- ***enseñanza asistida computarizada de lenguas***, que deben tener capacidad de análisis sintáctico para plantear y corregir ejercicios de gramática y composición

Departamento de Estudios Hispánicos
Recinto Universitario de Mayagüez
Edificio Chardón Oficina 503
Call Box 9000
Mayagüez, Puerto Rico 00681-9263

Portada: Vasily Kandinsky: Pintura Azul (https://www.guggenheim.org/artwork/1943)
Traducción y adaptación de materiales de Linguistic Society of America (www.linguisticsociety.org/content/computers-and-languages) y Association for Computational Linguistics (https://www.aclweb.org/portal/).
diseño y realización de Hilton Alers-Valentín, iv-2017
hilton.alers@upr.edu

RECINTO UNIVERSITARIO DE MAYAGÜEZ
CAAM
UNIVERSIDAD DE PUERTO RICO

Secuencia curricular en Lingüística Computacional

www.uprm.edu/linguistica

Perspectivas, problemas y acercamientos en LC

La lingüística computacional estudia lenguas naturales como el español y el japonés en vez de lenguajes de programación como C++ o Java. Este campo tiene dos perspectivas:

- la **tecnológica**, para hacer posible que las computadoras sirvan como instrumentos para analizar y procesar lenguas naturales
- la **psicológica**, para entender, por analogía con las computadoras, cómo los humanos procesamos una lengua natural

Desde ambas perspectivas, un lingüista computacional tratará de desarrollar un conjunto de reglas y procedimientos para, por ejemplo, reconocer la estructura sintáctica de oraciones o resolver referencias pronominales. Uno de los problemas más significativos al procesar lenguas naturales es el problema de la **ambigüedad.** En

(1) Él vio al hombre en el parque con el telescopio.

no está claro si es él, el hombre o el parque el que tiene el telescopio. De igual manera, si el inspector de bomberos te dice

(2) Hay una pila de basura inflamable al lado de su bicicleta. Va a tener que botar*la*.

que interpretes el pronombre *la* como si se refiriera a la pila de basura o a la bicicleta tendrá serias repercusiones en la acción que tomes. Ambigüedades como esta son ubicuas en enunciados orales y en textos escritos. La mayoría de las ambigüedades escapan nuestra atención porque somos muy eficientes para resolverlas usando nuestro conocimiento del mundo y del contexto. Pero los sistemas computarizados no tienen mucho conocimiento del mundo ni son eficientes usando el contexto. Para resolver este problema de la ambigüedad, existen dos acercamientos posibles: el **basado en conocimiento** y el **estadístico**.

En el acercamiento basado en conocimiento, hay que codificar mucho conocimiento del mundo y desarrollar procedimientos para usarlo al determinar el significado del texto. Para el ejemplo (2), habría que codificar hechos sobre el valor relativo de la basura y una bicicleta, sobre la conexión cercana entre los conceptos de *basura* y *botar*, sobre la preocupación de un inspector de bomberos por cosas que sean inflamables y así por el estilo. La ventaja de este acercamiento es que se parece más a la manera en que los humanos procesamos el lenguaje, por lo que probablemente resulte más eficiente a la larga. La desventaja es que el esfuerzo requerido para codificar el conocimiento necesario es enorme y que los procedimientos aplicados para utilizar este conocimiento son muy ineficientes.

Para el acercamiento estadístico se requiere un gran corpus de datos anotados. Luego se escriben procedimientos que computen las resoluciones más probables para las ambigüedades según las palabras o clases de palabras y otras condiciones fácilmente determinadas. Por ejemplo, se podrían recopilar triples palabra-preposición-nombre y aprender que el triple ⟨*vio, con, telescopio*⟩ es más frecuente en el corpus que los triples ⟨*hombre, con, telescopio*⟩ y ⟨*parque, con, telescopio*⟩. La ventaja de este acercamiento es que, una vez anotado el corpus, el procedimiento es automático y relativamente eficiente. La desventaja es que los corpus anotados requeridos son muy costosos y ningún corpus puede contener las producciones infinitas (y a veces improbables) del lenguaje natural. Además, los métodos obtienen análisis equivocados cuando la interpretación correcta requiere conciencia de factores contextuales sutiles.

Secuencia curricular en Lingüística Computacional

Esta secuencia curricular ofrece una formación competitiva en las áreas esenciales de la teoría y la aplicación de la lingüística computacional (LC) y el procesamiento de lenguas naturales (PLN). El conocimiento de la teoría lingüística, particularmente la sintaxis y la semántica, es esencial para entender los principios universales y los parámetros de variación en las lenguas naturales, las propiedades y rasgos del lenguaje como facultad distintiva de la especie humana y las gramáticas como representaciones mentales de un sistema cognitivo computacional. El estudio de los fundamentos formales de la lingüística computacional provee las herramientas lógico-matemáticas necesarias para el análisis y evaluación de modelos computacionales de aprendizaje, conocimiento y procesamiento lingüístico basados en sistemas deterministas y no-deterministas, simbólicos y probabilísticos, a la vez que permite familiarizarse con herramientas en línea para procesamiento de lenguas naturales, tales como corpus anotados, analizadores estructurales y redes y ontologías semánticas. La capacidad para programar en lenguajes de computación procedimentales y declarativos y para manejar diferentes técnicas y formatos de representación, estructura, almacenamiento y recuperación de información es indispensable para desarrollar modelos computacionales de PLN.

La Secuencia curricular en Lingüística Computacional es una especialidad complementaria disponible para cualquier estudiante del RUM, tanto regular (graduado o subgraduado) como de mejoramiento profesional. Consta de 18 créditos:

Cursos requisitos (15 créditos)

- LING 5030 Introducción a la Sintaxis
- LING 5060 Semántica Composicional
- LING 5080 Lingüística Computacional
- LING 5090 Fund. Form. de la Teor. Ling. (o CIIC 3075 o ICOM 4075 Fundamentos de Computación)
- COMP 3075 o CIIC 4020 o ICOM 4035 Estructura de datos

Cursos electivos (3 créditos)

- LING 4015 Seminario de Lingüística
- LING 4040 Fonética articulatoria y acúst.
- LING 5040 Introducción a la fonología
- LING 5050 Teoría Morfológica
- LING 5120 Psicolingüística
- COMP 5015 o CIIC5015 o ICOM 5015 Inteligencia Artificial
- COMP 5045 o CIIC 5045 Lenguajes formales y Autómatas

Prerrequisito

- LING 4010 El lenguaje en la mente humana

Facultad regular y adjunta

Hilton Alers-Valentín *(coordinador). Gramática generativa, semántica formal*
Melvin González. *Sintaxis, pragmática*
Doris Martínez. *Análisis del discurso*
Alexandra Morales. *Adquisición de lenguas, psicolingüística*
Nayda Santiago *(INEL). Arquitectura y organización computacional*
J. Fernando Vega *(ICOM). Procesamiento de lenguas naturales, Inteligencia Artificial*
Bienvenido Vélez *(CIIC). Compiladores, estructura de lenguajes de programación*