ASIGNATURA DE MÁSTER:

LINGÜÍSTICA COMPUTACIONAL

Curso 2012/2013

(Código:24402743)

1.PRESENTACIÓN

Presentación del equipo docente: Jefe del Departamento de Lingüística General y Lingüística Computacional de la Real Academia Española: Lingüística Computacional.

Es doctor en Filosofía y Letras por la Universidad Autónoma de Madrid (1997), universidad en la que ha sido profesor de Lingüística General y de Lingüística Computacional. Actualmente, es Jefe del Departamento de Lingüística Computacional de la Real Academia Española, institución a la que está vinculado, bien como colaborador, bien como personal de plantilla, desde 1996. Asimismo, imparte un curso de doctorado sobre Lingüística Computacional en el programa de doctorado de Lingüística Teórica y sus Aplicaciones en el I. U. Ortega y Gasset, dependiente de la Universidad Complutense de Madrid. Trabaja desde hace 20 años en Lingüística Computacional, disciplina en la que se ha formado gracias a su participación en proyectos europeos dedicados a la traducción automática (EUROTRA, 1987-1992), corrección gramatical (GramCheck, 1994-1995) y lingüística de corpus (CRATER, 1994-2005). En los últimos años, ha centrado su actividad en el desarrollo de herramientas y recursos de calidad para el procesamiento lingüístico automático en español, en sus dimensiones sincrónica y diacrónica (CREA y CORDE, 1995-2000), así como en el desarrollo de técnicas de extracción de términos a partir de corpus especializados (CCT, 2003-2006).

Éste es un curso en el que se pretende abordar los problemas computacionales del procesamiento lingüístico, fundamentalmente desde un punto de vista teórico, aunque sin renunciar a que el alumno tenga una experiencia práctica en el uso de herramientas y recursos habituales en la Lingüística Computacional.

2.CONTEXTUALIZACIÓN

El curso está inicialmente destinado a licenciados en filologías con un conocimiento medio-alto de los niveles de descripción, problemas y métodos en gramática. No se supone, en cambio, un conocimiento avanzado de los problemas de computación ni tampoco del manejo de los ordenadores, más allá del conocimiento como usuarios de las aplicaciones informáticas habituales. Con todo, los objetivos del aprendizaje están al alcance de otros licenciados, toda vez que el resto de los cursos de esta titulación de máster obligarán asimismo al alumno con menor formación en gramática a adquirir los conocimientos básicos para el satisfactorio aprovechamiento de los contenidos del máster en que este curso se enmarca.

La Lingüística Computacional aporta a los estudios lingüísticos el grado de formalización necesario para poder considerarlos científicos. En efecto, buena parte de los trabajos en Lingüística teórica, incluso realizados en el marco de escuelas con un fuerte aparato teórico, se basan más en la argumentación que en la formalización y no resisten una implementación del modelo propuesto. La Lingüística Computacional se apoya en tales teorías lingüísticas para, con ayuda de los formalismos gramaticales adecuados, representar el conocimiento lingüístico como tal y dar solución a problemas concretos del procesamiento lingüístico. Estas soluciones, estén cognitivamente motivadas o sean simplemente soluciones de ingeniería, permiten a los humanos resolver problemas cotidianos relacionados con el procesamiento lingüístico, y contribuyen a que el alumno descubra un ámbito bien definido de aplicación de sus conocimientos.

La Lingüística Computacional es una disciplina reciente que se ha incorporado a los planes de estudio de las licenciaturas de Lingüística. Sin embargo, como otros aspectos de la formalización lingüística, ha estado ausente de los planes de estudio de Filología y Humanidades. Este hecho ha favorecido la incorporación de Ingenieros y licenciados en ramas técnicas con conocimientos de programación al mercado de trabajo de la Ingeniería Lingüística en detrimento de quienes tienen la lengua como objeto de estudio. Este ámbito de la Ingeniería Lingüística es, sin duda, uno de los que mayor capacidad de absorción de nuevos licenciados

tiene. Baste, para ello, con echar un vistazo a las empresas que desarrollan tecnologías lingüísticas asociadas a la Red.

3.REQUISITOS PREVIOS RECOMENDABLES

Se espera que alumno tenga un conocimiento medio de los tradicionales niveles de descripción de la gramática, sus problemas y métodos. Asimismo, el alumno deberá tener capacidad lectora en inglés, pues buena parte del material didáctico está escrito en esta lengua. Finalmente, para la realización de ejercicios y prácticas, el alumno debe disponer de un ordenador.

4.RESULTADOS DE APRENDIZAJE

El curso pretende que el estudiante conozca los hitos más relevantes de la Lingüística Computacional, los problemas que se plantea y las aplicaciones que se desarrollan en esta área de investigación. Pretende, asimismo, que se familiarice con la formalización de problemas lingüísticos y para ello profundice en el estudio (y en la experimentación) de los formalismos de representación del conocimiento lingüístico más utilizados en la actualidad, resolviendo problemas concretos en varios de los niveles de descripción gramaticales.

El programa del curso persigue los siguientes objetivos en relación con la adquisición de conocimientos:

1. Que el estudiante revise y comprenda los fundamentos, problemas y métodos de los modelos computacionales del lenguaje dominantes en la actualidad: el probabilístico y el simbólico, con especial hincapié en el segundo de ellos.

2. Que conozca someramente la historia de la Lingüística Computacional, así como las aplicaciones más reseñables del Procesamiento del lenguaje natural.

3. Que estudie y, por tanto, conozca en profundidad, los formalismos más usuales para la representación del conocimiento lingüístico en los niveles morfológico, sintáctico y semántico.

5.CONTENIDOS DE LA ASIGNATURA

Junto a una panorámica histórica y metodológica, se presentan un conjunto de formalismos gramaticales/computacionales y la forma en que ayudan a modelar problemas de los distintos niveles de la gramática, con especial hincapié en los problemas de la sintaxis.

Siendo un curso introductorio a la Lingüística Computacional, que pretende formar mínimamente a los alumnos en esta joven disciplina encuadrable dentro de la Lingüística Aplicada, los contenidos de la asignatura abordan, con diferente grado de detalle, los ámbitos de trabajo, problemas y métodos más relevantes.

Junto a la panorámica histórica y a la presentación de modelos, tipos de aducto para el procesamiento y problemas fundamentales del procesamiento de las lenguas naturales, el curso presenta un conjunto de formalismos (y su relación con las máquinas abstractas correspondientes) capaces de tratar problemas propios de distintos niveles de descripción de la gramática. El estudio de estos aspectos permitirá ver los distintos grados de complejidad formal necesarios para el tratamiento de problemas en los diferentes niveles gramaticales. De estos, por su importancia en los estudios teóricos del último medio siglo, el contenido del curso destaca la sintaxis, nivel en el que el alumno tendrá que realizar un trabajo de implementación gramatical.

Por la importancia que los corpus textuales han adquirido en los últimos años en la investigación no solamente lingüística o tecnológica sino, en general, en todas las humanidades, el curso termina con un capítulo sobre anotación lingüística de corpus.

El material didáctico se estructura en 9 capítulos:

1. Introducción: Historia, modelos, aplicaciones.
2. Tecnologías del habla y tecnologías del texto.
3. Problemas de procesamiento de las lenguas naturales.
4. Introducción a Prolog.
5. Identificación de unidades de análisis.
6. Fonología/Morfología de estados finitos.
7. Sintaxis: Jerarquía de gramáticas, técnicas de *parsing*, gramáticas de unificación (PC-PARSE).
8. Desambiguación semántica.
9. Anotación de corpus.

6.EQUIPO DOCENTE

- RICARDO MAIRAL USON
- RAFAEL RODRIGUEZ MARIN

7.METODOLOGÍA

El curso incluye dos tipos de materiales, unos preparados expresamente para el mismo y otros que introducirán al alumno en otras tantas áreas de investigación dentro de la disciplina o que complementan el primer tipo de materiales. El objetivo de ambos es la transmisión de conocimientos teóricos. Sin embargo, los materiales del primer tipo incluyen ejercicios prácticos y proponen problemas cuya resolución será responsabilidad del alumno. Se pretende con ello una participación activa del alumno en el proceso de aprendizaje y un fortalecimiento de los aspectos prácticos del mismo, que permitan un seguimiento de los mismos adecuado por parte del profesor. Asimismo, el trabajo final de implementación de una gramática reúne, de algún modo, una parte importante de los saberes que el curso pretende transmitir al alumno, de forma que su realización pasa por la asimilación y la puesta en práctica de buena parte de los conocimientos adquiridos en el curso, a la vez que permite al alumno desarrollar sus aspectos más creativos en la elección y resolución de un problema de sintaxis de las lenguas naturales.

El estudiante debe enfocar la asignatura como un proceso continuo y constante, al que debe dedicar, si es posible todos los días, un tiempo razonable. Este enfoque favorecerá la comprensión y la capacidad de reflexión sobre el material que se le entrega. De hecho, algunos materiales son relativamente densos, tanto en la exposición, como en el tema mismo que abordan, lo que requiere una lectura pausada de los mismos y no un intento de asimilación veloz y voraz de los contenidos.

En general, el material no se ha desarrollado para el estudio memorístico sino más bien para la reflexión y el desarrollo de la capacidad de análisis y el razonamiento crítico. A estos objetivos contribuirá un tratamiento pausado y constante del material.

Se recomienda vivamente la lectura del material en el orden que se sugiere, pues la presentación de conceptos y problemas es incremental, por lo que otros itinerarios pueden llevar a la incomprensión y, por ende, al desaliento del estudiante.

Se recomienda seguir las pautas habituales de estudio, de forma que el trabajo con los textos se reduzca a las tareas de:

a) identificación de la estructura lógica del texto, con objeto de ir anticipando las cuestiones que el capítulo plantea;

b) lectura(s) del texto, buscando las ideas principales, los conceptos nuevos, su definición y su relación con los conceptos anteriores, del texto o de toda la asignatura;

c) resumen de las ideas principales del texto, realizando una reflexión, en los términos propios del alumno de las aportaciones del mismo;

d) asimilación de estas ideas, intentando fijar los conceptos fundamentales, relacionándolos con los conceptos aprendidos anteriormente y reproduciéndolos mentalmente;

e) realización de los ejercicios y resolución de los problemas planteados en todos aquellos capítulos que los incluyen. Unos y otros deberán realizarse después de haber leído y asimilado todos los conceptos que plantea el capítulo correspondiente;

f) relectura del texto: esta segunda lectura permitirá al alumno asimilar la terminología del dominio. Efectivamente, solo tras haber comprendido y asimilado los conceptos nuevos, se está en condiciones de asegurar la adecuada asimilación terminológica, que es, por otra parte, una de las características del discurso vinculado al conocimiento de una materia.

Una vez terminado el estudio de un tema y realizados los ejercicios y resueltos los problemas (si procede), el estudiante estará en condiciones de plantear sus dudas al profesor de la asignatura y de hacerle llegar su solución a los problemas. El profesor atenderá las dudas y comentará su solución a los problemas. Se creará una lista de distribución que permita enviar los comentarios a aquellas dudas que puedan resultar de interés general. Como ya se ha mencionado, de uno de los capítulos, el estudiante deberá enviar al profesor un resumen de lectura, en el que se destaquen las ideas principales y se realice una presentación crítica de los mismas.

Cuando el estudiante inicie el estudio del material didáctico correspondiente al capítulo sobre sintaxis, decidirá, junto con el profesor, el tema de trabajo práctico de implementación que deberá realizar como parte de los requisitos de evaluación del curso. En todo momento, por medio de las tutorías, recibirá consejos y comentarios sobre su trabajo hasta la entrega del mismo al final del curso.

8.BIBLIOGRAFÍA BÁSICA

Comentarios y anexos:

Los comentarios se introducen durante el curso y conforme a lo consignado en la Guía de la asignatura.

Jurafsky, D., and J. H. Martin (2000). *Speech and language Processing*. New Jersey: Prentice Hall.

Allen, J. F. (1994). *Natural Language Understanding*, 2nd edition, Reading, MA: Addison-Wesley.

Moreno Sandoval, A. (1998). *Lingüística Computacional*. Madrid: Síntesis.

Además de la básica, se indica otra relación de obras que en la Guía de la asignatura

En todas las lecturas se indica, con un número entre corchetes, el capítulo del temario al que se refieren. Asimismo, aquellas referencias que llevan además un asterisco se incluyen en el CD de la documentación. La bibliografía, para facilitar la

relación con el temario de la asignatura, se indica en el orden de lectura.

l Sánchez León, F. (2008) “Panorama general e histórico, definición(es), aproximación, modelos”. [1, *]

l European Commission (1997?) “Language Engineering: Harnessing the power of language”. [1, *]

l Oficina del Español en la Sociedad de la Información (OESI) “Tecnologías del habla” [Introducción en línea disponible en: http://oesi.cervantes.es/TLTODOS/tecnologias_del_habla.htm] El alumno deberá navegar por las páginas de este curso sobre tecnologías del habla. [2]

l Llisterri, J. (2993) “Las tecnologías del habla: Entre la ingeniería y la lingüística”, en *Actas del Congreso Internacional "La Ciencia ante el Público. Cultura humanística y desarrollo científico y tecnológico"*. Universidad de Salamanca, 28-31 de octubre 2002. Salamanca: Instituto Universitario de Estudios de la Ciencia y la Tecnología. Edición en CD-ROM. pp. 44-67. [2, *]

l King, M. (1988?) “A General Introduction to Natural Language Processing”. [3, *]

l Sánchez León, F. (2008) “Introducción a Prolog” [4, *].

l Grefenstette, G. and P. Tapanainen (1994) “What is a word, What is a sentence? Problems of tokenization”, in *3rd International Conference on Computational Lexicography*, Budapest, pp. 79-87. [5, *]

l Sánchez León, F. (2008) “Autómatas de estados finitos”. [6, *]

l Sánchez León, F. (2008) “Fonología/Morfología de estados finitos”. [6, *]

l Karttunen, L., J-P. Chanod, G. Grefenstette and A. Schiller (1996) “Regular Expressions for Language Engineering”. *Natural Language Engineering*, 2(4): 305-238. [6, *]

l Sánchez León, F. (2008) “Gramáticas y lenguajes formales”. [7, *]

l Sánchez León, F. (2008) “Introducción a los formalismos gramaticales basados en unificación”. [7, *]

l Sánchez León, F. (2008) “*Parsing*: estrategias”. [7, *]
l McConnel, S. (2002) “PC-PATR Reference Manual: A unification based parser”. [7, *]

l Agirre, E. and Ph. Edmonds (2006) “Introduction”, in *Word Sense Disambiguation: Algorithms and Applications*, Springer. [8, *]

l Sánchez León, F. (2008) “Anotación morfosintáctica de corpus”. [9, *]

de lecturas obligatorias.

9.BIBLIOGRAFÍA COMPLEMENTARIA

Comentarios y anexos:

Se presenta únicamente la bibliografía complementaria relativa a los dos primeros ttemas, pues el resto se ha integrado en los capítulos correspondientes del material didáctico. Como con las lecturas obligatorias, se incluye un asterisco al final de aquellas referencias que se entregan en el CD del curso:

l Gómez Guinovart, J. (1998): “Fundamentos de Lingüística Computacional: bases teóricas, líneas de investigación y aplicaciones”. En J. Baró i Queralt y P. Cid Leal (eds.), *Anuari SOCADI de Documentació i Informació*, Barcelona: Societat Catalana de Documentació i Informació, 135-146. Interesante artículo introductorio en el que se presentan los grados de imbricación entre Lingüística e Informática y las distintas líneas de trabajo en función de esa imbricación. Con todo, téngase

en cuenta que han transcurrido 10 años desde su publicación y que, por tanto, algunas limitaciones tecnológicas que plantea el artículo han sido superadas. Incluye una sección de fuentes de información relacionadas con esta disciplina. [*]

l Jurafsky, D. y Martin, J. H. (2000): Capítulo 1 “Introduction”. Repaso general a los problemas, modelos y algoritmos para el procesamiento lingüístico, junto con un breve apunte histórico. [*]

l Automatic Language Processing Advisory Committee (1966): *Language and Machines. Computers in Translation and Linguistics*. El alumno interesado en conocer las razones que se adujeron para el descalabro de la Traducción Automática a mediados de los 60, puede encontrar el informe ALPAC en la red http://books.nap.edu/openbook.php?isbn=ARC000005

l Moreno Sandoval, A. (1998): *Lingüística computacional*. Síntesis, Madrid. Se trata de una introducción básica muy asequible, por este motivo se ha incluido en la bibliografía recomendada. No obstante, debe manejarse con cautela, pues, aunque el planteamiento de la obra es adecuado, contiene algunos errores de fondo. El alumno deberá contrastar con el profesor de la asignatura aquellos aspectos contradictorios con las ideas del resto de la documentación del curso.

10.RECURSOS DE APOYO AL ESTUDIO

Los programas y lenguajes de programación que se utilizarán para ejemplificar aspectos del curso se incluyen en el CD con el material didáctico. Debe hacerse notar que este no es un curso de programación para lingüistas y que, por tanto, será difícil, en la mayoría de los casos, usar algunos de los programas y lenguajes para algo más que probar en casa los ejemplos que se explican en la documentación del curso.

En la elección de programas y lenguajes se ha optado por aquellos que se encuentran en el dominio público (son gratuitos) y que tienen versiones bajo sistema operativo Windows (en cualquiera de sus versiones). Se ha supuesto, por tanto, que el alumno tiene alguna versión de este sistema operativo instalada en su ordenador. Todo el material, sin embargo, se ha desarrollado y probado bajo Linux, sistema operativo para el que, lógicamente, existen versiones de estos mismos programas.

Para instalar los programas incluidos en el CD que contiene el material didáctico, el alumno deberá seguir las instrucciones que se incluyen en cada caso. Todos los programas se encuentran en directorio software. Todos tienen un desinstalador, por lo que podrán eleiminarse del disco duro, si se desea, cuando termine el curso.

A título meramente indicativo, además de un programa de descompresión de archivos, el mencionado directorio contiene un intérprete de Perl (ActivePerl), un intérprete de Prolog (SWI-Prolog) y el conjunto de programas para el desarrollo de procesadores morfológicos y sintácticos con gramáticas de unificación PC-PARSE. Asimismo, se incluyen programas ejemplo, desarrollados expresamente para el curso, con los que verificar determinados aspectos prácticos del curso.

11.TUTORIZACIÓN Y SEGUIMIENTO

El alumno será asesorado a lo largo de todo el curso a través de tutorías, en las que se le ayudará a resolver cualquier duda relativa a la materia impartida. Con este fin, se establece el horario que se indica más adelante, durante el cual el alumno podrá consultar telefónicamente al profesor cualquier asunto que sea de su interés. Asimismo, el profesor contestará a la mayor brevedad posible cualquier consulta que le sea remitida por correo electrónico. De forma excepcional, las consultas por carta ordinaria deberán realizarse a la dirección indicada más abajo.

Horario de atención al alumno

Días de la semana: Martes y Jueves

Horas: 16:30 a 18:30 horas

partamen

Dirección postal: Departamento de Lingüística Computacional

Centro de Estudios de la Real Academia Española

c/ Serrano, 187-189

28002 Madrid

Teléfono: 91 745 55 35

Correo electrónico: fsanchez@rae.es

12.EVALUACIÓN DE LOS APRENDIZAJES

Trabajo práctico, junto con una evaluación continua de los progresos/resolución de problemas por parte del alumno y del desarrollo de su capacidad de análisis y razonamiento crítico en el estudio.

Para superar el curso, el estudiante deberá enviar al profesor todos los ejercicios que se proponen en el material didáctico. Se espera una solución razonada (dado que todos los ejercicios obligan a realizar un esfuerzo de reflexión y adaptación de ejemplos presentados en el material didáctico) en la que se documenten las decisiones adoptadas. Asimismo, tras comunicación con el profesor de la asignatura, éste le asignará uno de los artículos de lectura obligatoria para que, tras su estudio (quizá complementado con la lectura del material complementario), redacte un resumen crítico de las ideas fundamentales del artículo. Este resumen puede realizarse a modo de presentación en aplicaciones como PowerPoint y, en cualquier caso, deberá enviarse al profesor, como la solución a los ejercicios, por correo electrónico.

La calificación final, además de tener en cuenta el trabajo de los estudiantes durante el curso, del que el profesor tendrá información puntual por medio de los ejercicios y el resumen de lectura, se fundamentamentará en el trabajo de diseño, implementación y documentación de una gramática sintáctica de la lengua de elección del estudiante; en este trabajo se desarrollará una solución de ingeniería a un problema concreto del nivel sintáctico. Por su parte, la documentación de la gramática deberá tratar tanto los aspectos lingüísticos como computacional, y deberá acompañarse de comentarios sobre cobertura y sobregeneración. Las directrices para la realización del trabajo se enviarán junto con el material didáctico.

13.COLABORADORES DOCENTES

- FERNANDO SÁNCHEZ LEÓN