




**UNIVERSIDAD DE BUENOS AIRES**  
**FACULTAD DE CIENCIAS EXACTAS Y NATURALES**

1. DEPARTAMENTO: Computación
2. CUATRIMESTRE: Primero de 2003
3. ASIGNATURA: Introducción al Procesamiento del lenguaje natural
4. CARRERA: Licenciatura en Ciencias de la Computación
5. CARACTER DE LA MATERIA: Optativa
6. NUMERO DE CODIGO DE CARRERA: 18
7. NUMERO DE CODIGO DE MATERIA: C 595
8. PUNTAJE: 2
9. PLAN DE ESTUDIOS AÑO: (1987 y 1993)
10. DURACION DE LA MATRERIA: cuatrimestral
11. HORAS DE CLASE SEMANAL:
  - a) TEORICAS: 2 hs
  - b) LABORATORIO: hs
  - c) PRACTICAS: 2 hs
  - d) SEMINARIOS: hs
12. CARGA HORARIA TOTAL SEMANAL: 4 hs,
13. ASIGNATURAS CORRELATIVAS: Probabilidades y estadística; Teoría de lenguajes y autómatas
14. FORMA DE EVALUACION: 2 parciales, promocional.
15. PROGRAMA Y BIBLIOGRAFIA: Adjuntas a esta hoja.
16. DOCENTES: Alvarez, José Angel

Fecha: 15/11/2002

  
Alvarez, José Angel

  
Dr. Guillermo Durai.  
Director Adjunto  
Depto. de Computación  
F. C. E. N. U. B. A.

## Introducción al Procesamiento del lenguaje natural – Programa y Bibliografía

### Objetivos:

- Proporcionar los fundamentos lingüísticos utilizados en el procesamiento del lenguaje natural.
- Introducir en los conceptos fundamentales, métodos y técnicas de procesamiento del lenguaje natural.
- Realizar aplicaciones de los conocimientos adquiridos a diversas tareas de procesamiento de lenguaje natural.

Docentes: Alvarez, José Angel

Dedicación: 4 horas semanales

Correlatividades: Probabilidades y estadística; Teoría de lenguajes y autómatas

### Contenidos

#### Unidad 1

Introducción a la disciplina del procesamiento del lenguaje natural (PLN). Enfoques racionalistas versus empiristas. Dificultades principales del PLN.

#### Unidad 2

Elementos básicos de estadística. Probabilidades e inferencia estadística. Teoría de la información: entropía, entropía condicional, codificación e información mutua. Elementos de reconocimiento de patrones. Teoría bayesiana de decisión. Construcción de clasificadores. Métodos de estimación no supervisados. Árboles de clasificación y regresión.

#### Unidad 3

Conceptos básicos sobre la estructura del lenguaje. Morfología y categorías sintácticas léxicas y frasales. Sintaxis, semántica, pragmática y discurso. Trabajo elemental con corpora, principales procesos y técnicas.

#### Unidad 4

Análisis principales de nivel léxico. Colocaciones y modelos de n-gramas. Principales técnicas de estimación. Desambiguación de sentidos léxicos, principales técnicas. Adquisición léxica. Principales cuestiones y técnicas.

#### Unidad 5

Análisis principales de nivel gramatical. Modelos markovianos ocultos. Etiquetado (tagging), principales métodos. Gramáticas probabilísticas independientes del contexto. Parsing probabilístico.

#### Unidad 6

Estudio de algunas de las principales aplicaciones. Alineación estadística y traducción automática. Recuperación de Información. Categorización de textos. Extracción de información.

### Bibliografía

- Allen, James (1995). *Natural Language Understanding*, Redwood City, Benjamin/Cummings, 2a. ed.
- Appelt, Douglas E. e Israel, David J. "Introduction to Information Extraction Technology". IJCAI-99, 1999.
- Bagga, Amit. *Coreference, Cross-Document Coreference, and Information Extraction Methodologies*. Ph.D. Thesis, Dept. of Computer Science, Duke University. June 1998.

Dr. Guillermo Duran  
Director Adjunto  
Depto. de Computación  
ECCOM IIRA

- Breiman, L., Friedman, J. Olshen, R. and Stone, C. Classification and regression trees, Wadsworth and Brooks, Pacific Grove, CA. 1984.
- Califf, May Elaine. *Relational Learning Techniques for Natural Language Information Extraction*. Ph. D. Thesis, Faculty of the Graduate School, The University of Texas at Austin, 1998.
- Cardie, Claire. "Empirical Methods in Information Extraction". *AI Magazine*, 18 (4), 1997.
- Charniak, Eugene (1993). *Statistical Language Learning*, Cambridge, MIT Press.
- Craven, Mark; DiPasquo, Dan; Freitag, Dayne; McCallum, Andrew; Mitchell, Tom; Nigam, Kamal y Slattery, Seán. "Learning to construct knowledge bases from the World Wide Web". *Artificial Intelligence*, 118 (2000), pp. 69-113.
- Cunningham, Hamish; Maynard, Diana; Bontcheva, Kalina; Tablan, Valentin; Ursu, Cristian y Dimitrov, Marin. "Developing Language Processing Components with GATE (a User Guide)". University of Sheffield, 2002.
- Fosler-Lussier, Eric (1998). "Markov Models and Hidden Markov Models: A Brief Tutorial", International Computer Science Institute, TR-98-041.
- Grishman, Ralph. "Adaptive Information Extraction and Sublanguage Analysis". IJCAI01, 2001.
- Jurafsky, Daniel y Martin, James H. *Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Upper Saddle River, New Jersey, 2000.
- Manning, Christopher D. y Schütze, Hinrich. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, 1999.
- Pazienza, Maria Teresa (comp.). *Information Extraction*. Springer-Verlag, Nueva York, 1997.
- Pustejovsky, James. *The Generative Lexicon*. MIT Press, Cambridge, 1995.
- Rajman, M. y Besançon, R. "Text Mining: Natural Language Techniques and Text Mining Applications". *Proceedings of the seventh IFIP 2.6 Working Conference on Database Semantics (DS-7)*, Chapam & Hall IFIP Proceedings series, (1997) Oct 7-10.
- Ritchie, Graeme D. y otros (1992). *Computational Morphology*, Cambridge, MIT Press.
- Sag, Ivan A. y Wasow, Thomas. *Syntactic Theory: A Formal Introduction*. Stanford: CSLI Publications. 1999.
- Sells, Peter y otros (1991). *Foundational Issues in Natural Language Processing*, Cambridge, MIT Press.

  
Dr. Guillermo Durán  
Director Adjunto  
Depto. de Computación  
E C E Y N U I R A