

Materia: Modelos Formales no Transformacionales

Departamento:

Letras

Profesor:

Reynoso, Carlos

2°Cuatrimestre - 2016

Programa correspondiente a la carrera de Letras de la Facultad de Filosofía y Letras de la Universidad de Buenos Aires

Programas



UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE FILOSOFIA Y LETRAS

DEPARTAMENTO: LETRAS

MATERIA: MODELOS FORMALES NO TRANSFORMACIONALES

PROFESOR: CARLOS REYNOSO

CUATRIMESTRE: 2°

Aprobado por Resolución

AÑO: 2016

N° 2355/16

PROGRAMA N°: 0581

MARTA DE PALMA
Directora de Despacho y Archivo General





UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE FILOSOFÍA Y LETRAS
DEPARTAMENTO DE LETRAS
MATERIA: Modelos Formales No Transformacionales
PROFESOR: Carlos Reynoso
CUATRIMESTRE y AÑO 2º Cuatrimestre 2016
PROGRAMA N° 0581



1. Fundamentación y descripción

En la tradición de la lingüística formal, se le reconoce a la obra de Chomsky *Estructuras Sintácticas* (1957) un rol fundacional. Una de las principales tesis del modelo del '57 se refiere a la complejidad de los lenguajes naturales, proponiendo a la sintaxis como un componente autónomo que genera oraciones mediante reglas de estructura de frase – también conocidas como Gramática Independientes de Contexto (*Context Free Grammar*, CFG) – y la aplicación *a posteriori* de reglas transformacionales.

La irrupción de la Gramática Generativo-Transformacional del '57 marcó un cambio de paradigma en la investigación lingüística. Sin embargo, a partir de la aproximación formal a las propiedades de los lenguajes naturales, han surgido diversas propuestas superadoras de la postura chomskyana inicial – el mismo Chomsky habría de dejar de lado el componente transformacional en sus modelos posteriores del '81 (Principios y Parámetros), y '95 (Programa Minimalista).

El presente programa ha sido pensado para alumnos tanto de la carrera de Letras de la orientación en Lingüística como de la carrera de Filosofía con orientación en Lógica, y para todos aquellos que tengan un fuerte interés en el estudio de la estructura sintáctica y del lenguaje natural y, en especial, en los alcances de la utilización de los modelos formales para el abordaje del mismo.

2. Objetivos

El objetivo fundamental de este curso es centrarse en el relevamiento minucioso de las teorías gramaticales NO transformacionales que evolucionaron a partir de la primera gramática chomskyana y que representan una importante línea de investigación en Procesamiento de Lenguaje Natural o *Natural Language Processing* (NLP), poniendo especial énfasis en las propiedades formales, poder expresivo y limitaciones de cada modelo en el estado del arte. Asimismo, se busca que el alumno trabaje con los rudimentos de implementaciones de dichas gramáticas bajo la forma de *parsers* (*analizadores sintácticos*), demostrando aplicaciones prácticas en español de los modelos teóricos.

3. Contenidos



La materia será dictada en 4 unidades que delinean un preciso recorrido teórico a través de los Modelos Formales No Transformaciones, desde sus orígenes en la lingüística chomskyana hasta las propuestas más recientes que evidencian marcadas mejoras en el Procesamiento de Lenguaje Natural y que atestiguan un cambio de paradigma científico desde los modelos simbólicos a los enfoques estadísticos. Adicionalmente, el curso incluye una quinta unidad de reflexión integral acerca de las implicancias lingüísticas y epistemológicas de una teoría gramatical.

Unidad I: Bajo la égida de Chomsky

Se introducirán la obra fundacional de Chomsky y las nociones formales con las que se ha de trabajar más adelante. Puntualmente se expondrá la jerarquía chomskyana respecto de las gramáticas (desde las más simples modelos de autómatas a las Máquinas de Turing, pasando por las Gramáticas Independientes de Contexto CFG). A su vez se expondrán las limitaciones en el poder expresivo de cada una de las Gramáticas, recurriendo tanto a lenguajes formales como a lenguajes naturales: por ej. el problema de la recursividad en la gramáticas de autómatas y la concordancia básica sujeto-verbo y el orden variable de constituyentes en las CFGs.

- 1.- Gramáticas: reglas de estructura de frase y transformacionales
 - 1.1 Dependencia e independencia del contexto
 - 1.2 Transformaciones sintácticas
- 2.- La jerarquía de Chomsky respecto de las gramáticas formales
- 3.- Tipos de autómatas: autómata finito, autómata descendente, autómata lineal
- 4.- Lenguajes regulares y autómatas
- 5.- Gramática Independiente de Contexto (CFG). Ventajas y limitaciones

Unidad II: La superación del modelo chomskyano

Se desarrollarán modelos formales alternativos a la CFG que trabajan con predicados lógicos de primer orden y el operador lógico *Unificación*. En particular se presentará la Gramática de Cláusula Definida – *Definite Clause Grammar* (DCG). Se puntualizará su utilidad como la primera propuesta de superación de las limitaciones expresivas de las Gramáticas Independientes de Contexto, aunque también se marcarán los problemas detectados en este tipo de modelos: por ej. la concordancia sujeto-verbo y la subcategorización verbal en español.

- 1.- Gramática de Cláusula Definida (DCG)
 - 1.1 Lógica de predicados de primer orden
 - 1.2 El operador lógico Unificación
- 2.- Ventajas sobre CFG. Aplicaciones prácticas en lenguaje natural
- 3.- Poder expresivo y limitaciones

Unidad III: Modelos optimizados fuera del área de injerencia chomskyana



Se abordarán las gramáticas basadas en la estructura de *rasgos (features)*. El gran salto cualitativo que representan estas gramáticas respecto del paradigma chomskyano es que intentan resolver los problemas tradicionales de los lenguajes naturales -como constituyentes discontinuos u orden variable en español- a través de operaciones de unificación de matrices de *features*, propiedades de herencia y restricciones de linealidad. Específicamente se trabajará con la gramática denominada Gramática de Estructura de Frase gobernada por Núcleos *Head-driven Phrase Structure Grammar-HPSG*, desarrollada por Ivan Sag.

- 1.- El concepto de *rasgo (feature)*
- 2.- Matriz de *feature*
- 3.- Implementación: herencia, jerarquías de tipos y restricciones de linealidad
- 4.- Head-driven Phrase Structure Grammar-HPSG Aplicaciones en español

Unidad IV: Paradigmas dominantes en el campo de investigación actual

La última unidad presentará el paradigma de investigación dominante en la actualidad: los enfoques estadísticos. El paradigma estadístico aplica diversas técnicas probabilísticas a grandes *corpora* de entrenamiento, con vistas a inducir categorías y fenómenos específicos del lenguaje natural a partir de la detección de patrones estadísticamente significativos en la *tabula rasa* que constituyen los *corpora*. Sin embargo, el paradigma estadístico es más que una mera aplicación de técnicas y modelización matemática: estos enfoques aportan evidencia de plausibilidad psicolingüística a un renovado debate acerca de la naturaleza misma del lenguaje.

- 1.- Principios fundantes del paradigma estadístico
 - 1.1 Cadenas de Markov y la equiparación de probabilidad y gramaticalidad
 - 1.2 Información distribucional e información mutua
- 2.- El debate epistemológico en torno al Argumento de la Pobreza de los Estímulos (APS).
La investigación actual sobre *corpus*
 - 2.1 Anotación morfosintáctica de *corpus*
 - 2.2 Técnicas de clustering e inducción de fenómenos sintácticos
 - 2.3 Aprendizaje automático o *machine learning*

Unidad V: El sustento epistemológico de las gramáticas no transformacionales

Esta unidad estará dedicada a establecer los alcances y los límites de las gramáticas no transformacionales desde un enfoque epistemológico y en el marco de la teoría lingüística. Se pondrá especial énfasis en los problemas conceptuales que subyacen a estas propuestas, así como en los conceptos teóricos y compromisos metodológicos que las sustentan.

- 1.- ¿Qué es una gramática como modelo formal?
 - 1.1 El rol preponderante de la sintaxis.
 - 1.2 Interacción entre los componentes sintáctico y semántico



2.- Evidencia psicolingüística para los modelos formales no transformacionales. Parsing y Procesamiento de Lenguaje Natural dentro de la Lingüística Computacional

4. Bibliografía específica (subrayada la bibliografía obligatoria)

Unidad I

- Chomsky, N. 1957. Estructuras sintácticas. México, SigloXXI. (hay diversas ediciones en español)
- Eguren, L. y Fernández Soriano, O. 2004. Introducción a una sintaxis minimalista. Madrid, Gredos. Cap.1
- Grune, D. y Jacobs, C. *Parsing Techniques. A practical guide.* Amsterdam, Springer, 2008
- Jurafsky, D y Martin, J. 2000. Speech and language processing. New Jersey. Prentice-Hall. caps. 2 y 9
- Partee, B., Meulen, A. y Wall, R. 1993. *Mathematical Methods in Linguistics.* Dordrecht: Kluwer Academic Publishers.
- Savich, W. Bach, E., Marsh, W. y Safran-Naveh, G. 1987 *The Formal Complexity of Natural Languages* Dordrecht: Kluwer Academic Publishers.
- Sag, I. And Wasow, Th. 1999. *Syntactic Theory. A Formal Introduction.* Stanford, California: CSLI Publications. Apéndice B.
- Sandoval Moreno, Antonio. 1998. Lingüística Computacional. Madrid. Editorial Síntesis
- Hierro S. Pescador, José. 1980. Principios de Filosofía del Lenguaje I. Teoría de los signos y de la Gramática. Madrid. Alianza Editorial cap.4

Unidades II y III

- Alsina, A. 1996. *The Role of Argument Structure in Grammar. Evidence from Romance.* Stanford, California: CSLI Publications.
- Jurafsky, D y Martin, J. 2000. Speech and Language Processing. New Jersey. Prentice-Hall. cap. 11
- Radford, A. 1997. *Syntactic Theory and The Structure of English. A Minimalist Approach.* Cambridge: Cambridge University Press.
- Radford, A. 1997. *Syntax. A Minimalist Introduction.* Cambridge: Cambridge University Press.
- Pollard, C. y Sag, I. 1987. *Information-Based Syntax and Semantics.* Stanford, CA: CSLI Publications.
- Pollard, C. y Sag, I. 1994. *Head-Driven Phrase Structure Grammar.* Stanford, CA: CSLI Publications.
- Sag, I. y Wasow, T. 1999. *Syntactic Theory. A Formal Introduction.* Stanford, California: CSLI Publications.
- Sells, P. 1989 *Teorías sintácticas actuales (GB, GPSG, LFG)* Barcelona, Ed. Teide.



- Russell S. y Norvig P. 1995. *Inteligencia Artificial: un enfoque moderno*. New Jersey. Prentice Hall caps. 22 y 23
- Sandoval Moreno, Antonio. 2001. *Gramáticas de Unificación y Rasgos*. Madrid .Ed. Antonio Machado
- Shieber, S. 1986. *An Introduction to Unification-Based Approaches to Grammar*. Stanford, CA: CSLI Publications.

Unidad IV

- Abney, S. 1996. "Statistical methods and linguistics". En: Klavans J. y Resnik, P. (eds.), *The balancing act*. The MIT Press, Cambridge, MA.
- Balbachan, F. 2014. *Técnicas de clustering para inducción de categorías sintácticas en un corpus de español*. Tesis de doctorado (UBA-FFyL).
- Balbachan, F. y Dell'Era, D. 2010. "Inducción de constituyentes sintácticos en español con técnicas de clustering y filtrado por información mutua". En *Linguamática* n°2, pp.39-57
- Chater, N. y Manning, C. 2006. "Probabilistic models of language processing and acquisition". En *TRENDS in Cognitive Sciences*, vol.10 n°7, pp.335-344
- Clark, A. 2001. *Unsupervised language acquisition: theory and practice*. Sussex. School of Cognitive and Computing Sciences, University of Sussex Press.
- Jurafsky, D y Martin, J. 2000. *Speech and Language Processing*. New Jersey. Prentice-Hall. caps. 1, 10, 15, 21
- Manning, C. y Schütze H. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press. Cambridge (Massachusetts). cap. 1
- Redington M. et al. 1998. "Distributional information: a powerful cue for acquiring syntactic categories". En *Cognitive Science* Vol 22 (4) 1998. pp. 425-469

Unidad V

- Chomsky, N. 1985. *El conocimiento del lenguaje*. Versión española de Eduardo Bustos Guadaño. Madrid: Alianza Editorial, 1989. Capítulos 1 y 2.
- Peacocke, Ch. 1989. "When is a grammar psychologically real?" en George, A. (comp.) *Reflections on Chomsky*. Cambridge: Basil Blackwell.
- Pinker, S. 1994. *El instinto del lenguaje*. Versión española de José Manuel Igoa González. Madrid: Alianza, 1995.
- Savich, W. Bach, E., Marsh, W. Y Safran-Naveh, G. 1987. *The Formal Complexity of Natural Languages* Dordrecht: Kluwer Academic Publishers.
- Shieber, S. 1985. "Evidence against the context-freeness of natural language". En *Linguistics and Philosophy* (8) pp.333-343
- Soames, S. 1984. "Linguistics and Psychology" en *Linguistics and Philosophy* 7.
- Soames, S. 1985. "Semantics and Psychology" en Katz, J. (comp.) *The Philosophy of Linguistics*. Cambridge: Oxford University Press.



5. Bibliografía general

- **Balbachan, F. (editor) Dell'Era, D., Berros J., Cruces, M. Haimovici, S. 2009. Cuaderno de ejercicios y apuntes. Serie Fichas de Cátedra. (FFyL-UBA)**
- Bolshakov, I. y Gelbukh, A. 2004. *Computational Linguistics: model, resources, applications*. México. UNAM
- Grune, D. y Jacobs, C. 2007. *Parsing techniques: a practical guide*. Amsterdam. Springer

6. Carga horaria

La materia estará dividida en clases teóricas y prácticas. En las primeras, se realizarán exposiciones acerca de los contenidos expuestos en el ítem anterior. En las segundas, se resolverán ejercicios relacionados con la parte teórica. Oportunamente se repartirá una guía de ejercicios para cada trabajo práctico.

Total de horas semanales: 6 horas

Total de horas cuatrimestrales: 96 horas

7. Condiciones de regularidad y régimen de promoción

La materia puede cursarse bajo el régimen de promoción directa. Se debe asistir a un 80% de las clases teóricas y a un 80 % de clases de trabajos prácticos y teórico-prácticos. Se deben aprobar dos parciales y un trabajo escrito con nota promedio igual o superior a siete (7) puntos.

Los alumnos que no hayan satisfecho los requisitos establecidos para la promoción directa, pero que hayan asistido al 75% de las clases de trabajos prácticos y aprobado los trabajos con un promedio no menor a 4 puntos son alumnos regulares y podrán presentarse en tal condición en la mesa general de exámenes.

Composición de la nota final:

4 mini trabajos prácticos grupales (10% cada uno = 40%)

2 parciales (25% cada uno = 50%)

Participación en clase y en foros del campus virtual (10%)

8. Planificación detallada de clases teóricas y trabajos prácticos



Semana	Clase #	Tema	Trabajos Prácticos	Bibliografía mínima para la clase
1	1-2	Introducción: Por qué MFNT ?	-----	1) Sandoval, Lingüística Computacional. 1998. cap. 1 y 2
2	3-4	Modelo Chomsky 1957. Jerarquía de Lenguajes Formales	Complementa teórico	1) Chomsky, Estructuras Sintacticas. 1957. caps. 3, 4, 5, 7 2) Hierro Pescador, Principios de Filosofía del Lenguaje. 1. Teoría de los Signos, de la gramática. Alianza Universidad 1980. sección 4.6 3) Eguren y Fernández Soriano. cap.1
3	5-6	Gramáticas Regulares. Automatas y máquinas de estado finitos. Expresiones regulares	Mini TP 1 Automatas	1) Sandoval, Lingüística Computacional. 1998. cap. 3 2) Jurafsky & Martin, Speech & Language Processing cap.2
4	7-8	Gramáticas Independientes de Contexto (Context Free Grammar CFG). Ventajas y limitaciones con ejemplos de Español e Inglés	Mini TP 2 CFG	1) Sandoval, Lingüística Computacional. 1998. sección 4.1 2) Jurafsky & Martin, Speech & Language Processing cap.9
5	9-10	Parsers Top-Down y Bottom-up (Bread first y depth first)		1) Sandoval, Lingüística Computacional. 1998. sección 4.3 2) Jurafsky & Martin, Speech & Language Processing cap.10
6	11-12	Chart parser. Optimización. Limitaciones de Parsers en general con gramáticas CFG para el Español y el Inglés. Repaso para parcial	Complementa teórico	1) Russell S. y Norvig P. Artificial Intelligence, cap.23
7	13-14	Examen Parcial 1	Repaso	-----
8	15-16	Gramática de Cláusula Definida (Defined Clause Grammar DCG). Ventajas y limitaciones con ejemplos del Español y el Inglés	Mini TP 3 DCG	1) Sandoval, Gramáticas de Unificación y rasgos, 2001 cap.1,2,3,4 2) Russel & Norvig, Artificial Intelligence, 1995. cap. 22
9	17-18	Concepto de rasgo o feature para Head-driven Phrase Structure Grammar (HPSG). Herencia, jerarquías de tipos y restricciones de linealidad en rasgos.	Complementa teóricos	1) Sandoval, Gramáticas de Unificación y rasgos 2001. cap.6
10	19-20	Gramática de Estructura de Frase gobernada por Núcleos Head-driven Phrase Structure Grammar-HPSG	Complementa teóricos	1) Sandoval, Gramáticas de Unificación y rasgos 2001. cap.7,8,9 2) Jurafsky & Martin, Speech & Language Processing cap.11
11	21-22	Qué es Procesamiento de Lenguaje Natural (Natural Language Processing NLP) ? Formalización de otros fenómenos en NLP (más allá del parsing).	Complementa teóricos	1) Sandoval, Lingüística Computacional. 1998. cap.5 2) Jurafsky & Martin, Speech & Language Processing caps.15 y 21
12	23-24	Relación con otros paradigmas (enfoque estadístico y conexionismo) "Clustering y Parsers"	Mini TP 4 clustering	1) Sandoval, Lingüística Computacional. 1998. 6.1 y 6.2 2) Manning & Schultze, Foundations on statistical NLP cap.1 3) Balbachan. 2014
13	25-26	Repaso para examen 2 y discusión general de la materia	Complementa teóricos	1) Sandoval, Gramáticas de Unificación y rasgos, 2001. cap.10
14	27-28	Examen Parcial 2	Repaso	-----


MIGUEL VEDDA
 DIRECTOR
 DEPARTAMENTO DE LETRAS



Dr. Carlos Reynoso
 Legajo 93.862 (FFyL-UBA)
 Profesor adjunto
 con asignación de funciones a cargo de la cátedra