Simposio: Nuevas aplicaciones de la teoría de lenguajes formales a la lingüística

Gemma Bel Enguix y M. Dolores Jiménez López

Univesitat Rovira i Virgili, Grupo de Investigación en Lingüística Matemática
Pl. Imperial Tarraco, 1, 43005 Tarragona
gemma.bel@urv.cat mariadolores.jimenez@urv.cat

Resumen

El simposio que proponemos tiene como objetivo principal reunir una serie de comunicaciones de distintos investigadores que, procediendo de ámbitos de investigación muy variados, tienen en común la aplicación de la teoría de lenguajes formales a distintos aspectos del análisis, descripción y generación de las lenguas naturales.

Palabras clave: Teoría de lenguajes formales, lingüística.

Abstract

The aim of this workshop is to bring together researchers from different areas that have in common the use of formal language theory to approach different aspects of the analysis, description and generation of natural language.

Keywords: Formal language theory, linguistics.

Resum

El simposi que proposem té com a principal objectiu reunir una sèrie de comunicacions de diversos investigadors que, pertanyent a àmbits de recerca molt variats, tenen en comú l'aplicació de la teoria de llenguatges formals a diversos aspectes de l'anàlisi, descripció i generació de les llengües naturals.

Paraules clau: Teoria de llenguatges formals, lingüística

Tabla de contenidos

1. Introducción
2. Teoría de lenguajes formales
3. Teoría de lenguajes formales y lingüística
4. Comunicaciones del simposio
5. Referencias bibliográficas

1. Introducción

Vivimos inmersos en la llamada sociedad de la información en la que hay una gran necesidad de disponer de una amplia tecnología lingüística para la gestión de los ítems comunicativos. Mientras los ordenadores puedan comunicarse solamente por medio de lenguajes artificiales diseñados específicamente para ellos, la comprensión hombre-máquina no se podrá llevar a cabo de forma eficiente. Lo natural sería permitir al usuario dirigirse al ordenador en su propia lengua. Es aquí donde aparecen las llamadas *tecnologías del lenguaje*, entendidas como la aplicación de los conocimientos sobre la lengua al desarrollo de sistemas informáticos que puedan reconocer y generar lenguaje humano. La Unión Europea ha puesto de manifiesto el papel fundamental de este ámbito de estudio y cuenta con un importante programa de investigación en el campo de las tecnologías del lenguaje que ha evolucionado rápidamente en los últimos años.

Las tecnologías del lenguaje se proponen desarrollar sistemas informáticos que puedan reconocer y generar lenguaje. Por tanto, estas tecnologías necesitan modelos formales y

generales que capten la estructura del lenguaje natural y que sean eficaces desde el punto de vista computacional. Creemos que la teoría de lenguajes formales puede ofrecer las herramientas matemáticas necesarias para la definición de mecanismos altamente formales y computacionalmente válidos. Por ello proponemos este simposio, en el que se presentan distintas aplicaciones de la teoría de lenguajes formales al ámbito de la descripción, análisis y procesamiento del lenguaje natural.

Creemos que la teoría del lenguajes formales –sobre todo en su vertiente no-clásica, esto es los modelos formulados en los últimos años— presenta características que resultan muy útiles a la hora de analizar, describir y explicar cualquier problema lingüístico.

2. Teoría de lenguajes formales

En los años cincuenta, con las primeras propuestas de Chomsky, nace un nuevo modelo para el análisis de lenguaje y de las lenguas naturales. Este nuevo modelo lingüístico, el generativo, parece poder explicar mediante determinados formalismos un aspecto clave de las lenguas naturales: la *recursividad*. Con la propuesta formulada por Chomsky sobre el carácter formal y generativo del lenguaje, nace un ámbito de investigación: el de los *lenguajes formales*. Este nuevo ámbito de investigación, que es deudor de la lingüística por su filiación, está sólidamente asentado en la teoría de sistemas formales y en la lingüística algebraica, entendida como disciplina que maneja herramientas matemáticas para la descripción de fenómenos lingüísticamente reseñables, ya sea desde el punto de vista de los lenguajes naturales, ya sea desde la perspectiva de los lenguajes formales (Ortega i Robert 1993).

Todo lenguaje puede ser visto como un conjunto de frases. Una frase es una cadena finita compuesta de elementos extraídos de un vocabulario. La teoría de lenguajes formales se ocupa de la especificación sintáctica de los lenguajes, dejando de lado cualquier consideración semántica. Para lenguajes finitos, una posible especificación sintáctica consiste en dar la lista de las frases de que se compone. Para lenguajes infinitos, tal lista no es posible. Por ello, la principal tarea de la teoría de lenguajes formales es la especificación finitaria de lenguajes infinitos (Martín Vide 1994). En la teoría clásica de lenguajes formales, la mayoría de estas especificaciones son casos especiales de la noción de sistema de reescritura.

Un lenguaje formal es un conjunto arbitrario de cadenas (palabras) sobre un alfabeto (finito o infinito) V (llamado también vocabulario o diccionario). Una gramática formal es un mecanismo finito por medio del cual podemos generar los elementos de un lenguaje. En teoría de lenguajes formales se distinguen, fundamentalmente, dos tipos de mecanismos: 1) las *gramáticas*, que son dispositivos generadores; y 2) los *autómatas* que son dispositivos reconocedores.

La *teoría de autómatas*, también llamada *teoría algebraica de máquinas*, permite estudiar de modo sistemático las máquinas que realizan un procesamiento de la información y que actúan de manera discreta, esto es, la información se supone codificada a partir de un conjunto finito de símbolos que el autómata trata secuencialmente (Miquel i Vergés 1993). La teoría de autómatas está íntimamente relacionada con la teoría de lenguajes formales, en la medida en que los conjuntos de datos y de resultados de un sistema computacional cualesquiera pueden ser

considerados como lenguajes. Además, podemos hacer equivaler clases de modelos de computación con clases de sistemas de especificación. Por tanto, la teoría de lenguajes formales y la teoría de autómatas, que constituyen los fundamentos de las ciencias de la computación, están inseparablemente unidas.

Según (Martín Vide 1996), la teoría clásica de gramáticas formales ha hecho un uso casi exclusivo de los procedimientos de reescritura. A partir de la jerarquía de Chomsky, se han desarrollado diversas técnicas generativas, todas ellas más o menos próximas a las herramientas originales. Esto puede llevar a creer que la reescritura es inevitable en la teoría de lenguajes formales, en la teoría de autómatas, en la teoría de algoritmos o en las ciencias de la computación en general. Sin embargo, en los últimos años han aparecido, dentro de la teoría de lenguajes formales, herramientas generativas que no hacen uso de la reescritura y que tienen la misma capacidad generativa que los dispositivos clásicos. En muchas ocasiones, estas nuevas herramientas parecen más naturales que la reescritura y, por tanto, pueden ser más adecuadas para la descripción del lenguaje natural. Por este motivo, en este simposio, queremos considerar aplicaciones lingüísticas no solo de la teoría de lenguajes formales clásica sino también de los modelos y herramientas propuestos en los últimos años, esto es, de la llamada teoría de lenguajes formales no-clásica.

3. Teoría de lenguajes formales y lingüística

Nuestra propuesta de simposio no presenta una idea completamente nueva, sino que pretende continuar con una tradición de intercambio de métodos entre las ciencias de la computación y los lenguajes naturales. Computación, lingüística y teoría de lenguajes formales se han influido mutuamente durante años Las lenguas naturales han servido de modelo a la computación, y los lenguajes formales han sido usados como modelos en el estudio de las lenguas naturales. Tanto es así que, como hemos dicho, la teoría de lenguajes formales nació a mediados del siglo XX como una herramienta para describir la sintaxis de las lenguas naturales. A partir de 1964, esta teoría se desarrolló como una rama independiente de la lingüística, con problemas, técnicas y resultados específicos. Desde entonces ha desempeñado un importante papel en el ámbito de la computación.

Si las herramientas utilizadas en el estudio de las lenguas naturales han sido adecuadas para los lenguajes de programación, parece razonable pensar que los métodos desarrollados para el estudio de los lenguajes formales durante los últimos años puedan resultar útiles a la hora de dar cuenta de las lenguas naturales. Por este motivo, pensamos que las nuevas aplicaciones de la teoría de lenguajes formales a la descripción del lenguaje natural constituye un área de investigación interesante que puede arrojar buenos resultados en lingüística mediante la reformulación en la manera de describir la estructura de las lenguas y mediante la definición de modelos formales para su manipulación que puedan ser útiles en cualquier ámbito de la inteligencia artificial que implique el procesamiento del lenguaje natural.

Los modelos propuestos en los últimos años en el ámbito de la teoría de lenguajes formales presentan muchas ventajas respecto a los modelos clásicos. Modularidad, paralelismo, interacción, distribución, motivación biológica, etc. están en la base de muchas de las teorías que han surgido últimamente. Todas estas características apoyan la adecuación de estos modelos en el estudio de las lenguas naturales. Además, el hecho de que, la mayor parte de las veces, la investigación en teoría de lenguajes formales se

haya centrado básicamente en los aspectos teóricos de los modelos, hacen de la idea de aplicación de estos mecanismos a la lingüística un área de investigación innovadora que puede arrojar resultados interesantes y relevantes tanto en el ámbito de la teoría de lenguajes formales como en el campo del procesamiento del lenguaje natural.

4. Comunicaciones del simposio

El simposio que presentamos incluye cuatro comunicaciones en las que se propone la aplicación de distintos modelos formales a diferentes cuestiones lingüísticas. En concreto, los artículos que se presentan se ocupan de adquisición del lenguaje, formalización del diálogo, gramática y traducción automática. Los cuatro artículos tienen en común la utilización de herramientas procedentes de la teoría de lenguajes formales para abordar las diferentes cuestiones lingüísticas tratadas.

En la primera de las comunicaciones, la Dra. Becerra-Bonache propone la aplicación de la teoría de la inferencia gramatical a cuestiones relacionadas con la adquisición del lenguaje. El trabajo propuesto por la autora es claramente interdisciplinar. En él se combinan ideas provenientes de distintas áreas de conocimiento (lingüística, ciencia cognitiva y ciencias de la computación) con el objetivo de buscar un modelo que sea capaz de dar cuenta de la adquisición del lenguaje. Teniendo en cuenta algunas de las limitaciones que presentan los estudios de inferencia gramatical cuando se intenta aplicarlos al ámbito de la lingüística, la Dra. Becerra-Bonache propone un nuevo modelo en el que se proponen algunas ideas claramente innovadoras. En concreto, la autora propone dos objetivos básicos: 1) centrar los estudios de inferencia gramatical en clases de lenguajes relevantes desde el punto de vista lingüístico; y 2) considerar un nuevo modelo de aprendizaje en el área de la inferencia gramatical en el que se combine la disponibilidad de datos positivos (frases gramaticalmente correctas) y correcciones.

La segunda comunicación lleva por título "Definición de una Jerarquía de Clases de Protocolos de Diálogos" y pretende ser una contribución a la teoría del diálogo. La Sra. Grando define dos marcos formales inspirados en la teoría de lenguajes formales para la simulación de diálogos como sistemas de transiciones de estados finitos. En el primer marco formal, los agentes disponen de una memoria compartida que se limita a una pila que almacena locuciones. Esta pila parece ser adecuada para simular diálogos guiados por metas. La pila puede almacenar durante el diálogo las metas no conseguidas. El tope de la pila corresponde a la última meta a lograr, que es removida de la pila cuando es alcanzada. Este marco formal no puede simular todos los diálogos basados en la noción de semántica social. El principio subyacente en todo sistema basado en la semántica social es que cuando los participantes en el diálogo hablan, están dando a conocer públicamente su conocimiento y adquiriendo compromisos. La verdad de lo expresado por un hablante, en general, no puede ser verificado, pero al menos la consistencia de su discurso puede ser corroborada a través de los compromisos sociales que ha adquirido. Con el propósito de dotar al marco definido del poder expresivo necesario para simular diálogos basados en la noción de semántica social, la Sra. Grando propone la extensión de la pila de locuciones a una cadena de símbolos sobre un alfabeto finito.

El Sr. Perekrestenko propone una comunicación titulada "*Extending Tree-adjoining Grammars and Minimalist Grammars with unbounded scrambling: an overview of the problem area*". El autor parte de la idea de que, en su versión estándar, las gramáticas minimalistas son débilmente equivalentes a las gramáticas de adjunción de árboles de

conjunto local multicomponente y que la tractabilidad computacional de estas gramáticas se consigue por medio de la llamada condición del movimiento más corto (*shortest-move condition*, SMC), condición que prohíbe el desplazamiento competitivo de subárboles y que imposibilita la descripción generalizada del desplazamiento opcional competitivo ilimitado de constituyentes sintácticos: '*scrambling*'. Para solucionar este problema, el Sr. Perekrestenko propone una serie de modificaciones a las gramáticas minimalistas con scrambling ilimitado y con barreras y estudia sus consecuencias para la complejidad del problema de reconocimiento. Asimismo, se ocupa de cómo se pueden expresar en las gramáticas minimalistas las restricciones imponibles al proceso de derivación en las gramáticas de adjunción de árboles basadas en vectores no locales con dominancia y integridad que garantizan el reconocimiento en tiempo polinómico para los lenguajes generados y no perjudican la descripción generalizada del *scrambling*. Además, el autor analiza la posibilidad de modificar las gramáticas minimalistas de manera que se cumplan estas restricciones. Finalmente, se aborda la cuestión de las correspondencias de las modificaciones propuestas con las gramáticas de adjunción de árboles.

Por último, el Sr. Tîrnăucă propone un artículo, titulado *"Tree bimorphisms and their relevance in the theory of translations*", que se sitúa en el ámbito de la traducción automática. Teniendo en cuenta la idea de "interlingua" propuesta en los estudios de traducción automática en los años 60, así como la necesidad de una base matemática sólida en este ámbito, el autor propone una modificación del concepto original de "interlingua" con el objetivo de mejorar las bases matemáticas de la traducción automática. Utilizando herramientas procedentes de la teoría de lenguajes formales –los *tree transducers*— y herramientas lingüísticas –*synchronous grammars*—, el Sr. Tîrnăucă propone construir un lenguaje abstracto, una interlengua, para cada dos pares de lenguas naturales. El autor defiende la idea de que los *tree bimorphisms* pueden formalizar/modelar el concepto de "interlingua" y definir transformaciones de árbol.

5. Referencias bibliográficas

Martín Vide, C. (1994). "Gramáticas formales (y similares) para lingüistas". En C. Martín Vide, ed., *Lenguajes Naturales y Lenguajes Formales X*. Barcelona: PPU, pp. 71-91.

Martín Vide, C. (1996). "Computación Natural". En C. Martín Vide, ed., *Lenguajes Naturales y Lenguajes Formales XII*. Barcelona: PPU, pp. 121-127.

Miquel i Vergés, J. (1993). "Teoría de autómatas". En C. Martín Vide, ed., *Lenguajes Naturales y Lenguajes Formales IX*. Barcelona: PPU, pp. 115-131.

Ortega i Robert, R. (1993). "Teoría de gramáticas formales". En C. Martín Vide, ed., *Lenguajes Naturales y Lenguajes Formales IX*. Barcelona: PPU, pp. 99-113.

Partee, B.H., A. Meulen y R.E. Wall, R.E. (1993). *Mathematical Methods in Linguistics*. Dordrecht : Kluwer.

Păun, Gh. y A. Salomaa, eds. (1997). *New Trends in Formal Languages. Control, Cooperation, and Combinatorics.* Berlin: Springer.

Révész, G. (1983). *Introduction to Formal Languages*. New York: Dover Publications.

Rozenberg, G. y A. Salomaa, A., eds. (1997). *Handbook of Formal Languages*. Berlin: Springer.

Salomaa, A. (1973). *Formal Languages*. New York: Academic Press.