TESIS DOCTORALES

A Formal Approach to Spanish Syntax

Tesis presentada por
Jos Hallebeek
en la Facultad de Filosofía y Letras de la Universidad de Nijmegen (Países Bajos)
para obtener el Título de Doctor, en 1990.

Se publicará en versión inglesa en la editorial RODOPI, Amsterdam - Atlanta,GA, a principios de 1992.

En esta tesis se da cuenta de un proyecto de investigación llevado a cabo en el período que va desde agosto de 1985 hasta agosto de 1989 dentro del programa de Lingüística de Corpus de la Facultad de Letras de la Universidad de Nimega. El proyecto era conocido bajo el nombre de ASATE, formado por las siglas de Análisis Sintáctico Automatizado de Textos Españoles. Tenía como objetivo principal componer una gramática formal del español convertible en analizador morfosintáctico automatizado, un parser. Para cumplir con esta última condición utilizamos el formalismo de descripción de las Extended Affix Grammars (Gramáticas de Afijo Ampliadas). Estaba a nuestra disposición un parser generator diseñado para este formalismo por el departamento de Informática de la Universidad de Nijmegen. La gramática había de reunir una serie de requesitos que se resumen de esta forma:

1. describe el uso de la lengua, concretamente el español escrito contemporáneo;
2. contiene módulos para los sistemas de reglas morfológico y sintáctico; los aspectos semánticos y pragmáticos quedan fuera de consideración;
3. ofrece la descripción más consistente y completa posible dentro de los marcos que se acaban de indicar;
4. es capaz de analizar - previa conversión en parser - un fragmento de cualquier texto con un mínimo de intervenciones.

El estudio va dividido en cuatro capítulos. Los dos primeros son de carácter introductorio. El capítulo 1 tiene por objeto enmarcar el proyecto en la lingüística española. Ofrece una introducción al estado actual de la gramática descriptiva del español. En la composición de la gramática formal los estudios descriptivos juegan un papel importante. Constituyen la principal fuente de datos respecto al uso actual de la lengua. Y con la gramática que hemos compuesto creemos haber

contribuido al desarrollo de la gramática descriptiva del español. En la segunda parte del mismo capítulo se ofrece una visión de conjunto de las actividades de investigación dentro y fuera de España en el terreno del análisis gramatical automatizado del español. Se nota un interés creciente por la utilización del ordenador en la investigación de esta lengua. Hay una gran variedad de aplicaciones. Sin embargo, llama la atención la falta de intentos para realizar un sistema de análisis morfosintáctico automatizado completo de textos españoles sin lematizar. Según determinados linguistas montar un sistema de este tipo cuesta demasiado tiempo o resulta demasiado complicado. Casi la totalidad de las investigaciones llevadas a cabo o en curso de realización va encaminada a la obtención de resultados prácticos, siendo financiada por empresas comerciales o entidades europeas. Su objetivo primario es la producción de resultados de aplicación práctica. El método de investigación, así como los medios empleados no son siempre verificables del todo. Este no es el caso en la investigación sobre la que en este estudio informamos, dando cuenta paso a paso del método y de los datos empleados, así como de los resultados obtenidos.

El capítulo 2 contiene una descripción del formalismo de las Extended Affix Grammars, que proviene de la informática. Trata del origen, las aplicaciones y características del formalismo. La Extended Affix Grammar es un instrumento que se presta a la producción y al análisis tanto de lenguas artificiales como de lenguas naturales. De hecho, su campo de aplicación en la Informática es tan amplio que se la utiliza no sólo para la definición de lenguas de programación sino también como una lengua de programación. Se ha comprobado la adecuación del formalismo para la descripción de lenguas naturales desarrollando gramáticas (parciales) del inglés y del árabe estándar moderno.

El capítulo 3 constituye la parte fundamental del estudio. Es el más extenso, puesto que se da cuenta en él de la gramática formal, incluyendo el lexicón. Se pasa revista a todas las estructuras morfológicas, sintácticas y lexicales incluidas en las reglas. En todos los casos empezamos analizando cualquier estructura a base de estudios descriptivos pertinentes del español. Si es necesario se completan tales datos con intuiciones propias. Luego la estructura es vertida en reglas formales y encajada en la parte ya desarrollada de la gramática, de modo que forme una unidad coherente. La configuración de la gramática es modular. Sus partes se corresponden con las unidades gramaticales conocidas, tales como Noun Phrase (Sintagma Nominal), Adjective Phrase (Sintagma Adjetivo), Verb Phrase ((Sintagma Verbal), etc. Se dedican descripciones separadas a los fenómenos sintácticos que rebasan los límites de una unidad sintagmática. A este respecto cabe mencionar las construcciones comparativas, la coordinación y la subordinación. Las reglas formales correspondientes van incorporadas en las partes de la gramática con que se relacionan. En subcapítulos aparte se tratan además la extensa morfología del verbo español, la formación de palabras y la construcción del lexicón.

En la primera parte del capítulo 4 se discuten las pruebas a que se han sometido gramática y lexicón. Fueron realizadas en dos fases. La primera de ellas corría paralela a la construcción gradual de la gramática. El correcto funcionamiento de las reglas de cualquier estructura nueva era probado inmediatamente mediante el análisis de una serie de frases formuladas por el investigador. La segunda fase de pruebas se inició a raíz de terminarse la gramática, incluyendo el componente morfológico y el lexicón. Se utilizaron textos españoles originales. Entre ellos figuraban fragmentos de diferentes tipos de textos: un libro de viajes, una obra de teatro, un discurso político, una novela y dos artículos periodísticos. Este material se tomó del Nijmegen Corpus van Hedendaagse Spaanse Teksten (Corpus de Nimega de Textos Españoles Contemporáneos). Se trata de una colección de textos actuales disponible en soporte magnético, cuya fecha de publicación está entre 1975 y 1984. Tiene una extensión de 500.000 palabras y es otro producto salido del proyecto ASATE. El análisis automatizado se llevó a cabo en dos pasos: primero el análisis léxicomorfológico y luego el análisis sintáctico. Se obtuvieron los siguientes resultados. Tras añadir al lexicón las pocas radicales que faltaban, el parser léxicomorfológico produjo el análisis deseado de todas las palabras de los enunciados presentados. El parser sintáctico acertó al analizar el 70 por ciento de todos estos enunciados dentro del límite de tiempo CPU fijado en 1.000 segundos. Este límite se estableció a fin de evitar problemas de capacidad. Además de agregar algunas reglas nuevas a la gramática resultó necesario introducir algunas modificaciones en reglas ya existentes. Los enunciados cuyo análisis rebasó el límite fijado tenían una extensión y complejidad mayores de lo común. En general, los enunciados de más de veinte palabras daban lugar a este tipo de problemas de duración. Los resultados obtenidos de las pruebas constituían el material para la evaluación de la gramática. De esta evaluación trata la segunda parte del capítulo 4. En ella se comprueba que tanto el lexicón como la gramática cumplen con la exigencia de que la gramática sea capaz de analizar cualquier fragmento de texto con un número limitado de intervenciones. Resultó necesario aumentar el número de radicales que contiene el lexicón sólo en un 4 por ciento. El componente morfológico no necesitó ninguna modificación. Las limitaciones se encuentran en el componente sintáctico. Para enunciados de una extensión mayor de lo corriente, el análisis era muy largo, de modo que había que fijar un límite superior. Los enunciados cuyo análisis dio resultados muestran cierto número de análisis ambiguos, con un promedio de 5,8 para cada enunciado. Esta cifra no resulta alarmante y más bien era de esperar si se considera la falta de componentes semántico y pragmático de la gramática. Teniendo en cuenta las limitaciones que acaban de indicarse es justificada la conclusión de que la gramática desarrollada forma un instrumento adecuado para el análisis morfosintáctico de la mayor parte de cualquier corpus de textos españoles contemporáneos con un mínimo de intervenciones.