Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 335-336 recibido 19-04-2011 aceptado 23-05-2011

Desarrollo de Recursos para el Análisis Sintáctico Automático del Español: AVALON, una gramática formal y CSA, un corpus sintácticamente analizado*

Development of resources for the automatic syntactic analysis of Spanish: AVALON, a formal grammar, and CSA, a syntactically analysed corpus

M.ª Paula Santalla del Río
Universidad de Santiago de Compostela
Facultad de Filología, Avda. Burgo das Nacions, s/n, 15782 Santiago de Compostela
tfno.: 881811766, e-mail: mpaula.santalla@usc.es

Resumen: Presentación del proyecto de investigación DRASAE para el desarrollo de un corpus sintácticamente analizado y una gramática formal basada en datos reales extraídos de una base de datos sintácticos
Palabras clave: Gramática formal, corpus sintácticamente anotado

Abstract: Presentation of the research project DRASAE for the development of a treebank and formal grammar based on real data derived from a syntactic database
Keywords: Formal grammar, syntactically analysed corpus, treebank

1. El proyecto DRASAE[1]

El proyecto *DRASAE, Desarrollo de Recursos para el Análisis Sintáctico del Español: AVALON, una gramática formal, y CSA, un corpus sintácticamente anotado*, es un proyecto de investigación recientemente puesto en marcha por una sección del grupo de investigación *Gramática del español* perteneciente a la Universidad de Santiago de Compostela. De esa sección del grupo de investigación referido, forman parte los investigadores M.ª Paula Santalla del Río, que dirige el proyecto, Iria del Río Gayo, Guillermo Rojo y Susana Sotelo. El proyecto DRASAE, que ha recibido financiación para los próximos tres años (2011-2013) de la Dirección Xeral de Investigación, Desenvolvemento e Investigación de la Consellería de Economía e Industria de la Xunta de Galicia, se propone la mejora y ulterior desarrollo de un recurso lingüístico preexistente, una gramática formal del español (AVALON), así como la creación del entorno necesario y, simultáneamente, su generación efectiva, para la producción de uno nuevo, un corpus sintácticamente analizado.

* Agradecemos la ayuda brindada a esta investigación por la Dirección Xeral de Investigación, Desenvolvemento e Innovación de la Consellería de Economía e Industria de la Xunta de Galicia.

[1] URL: http://gramatica.usc.es/proxectos/drasae/

2. Desarrollo de AVALON

AVALON es una gramática formal en el formalismo AGFL[2], a partir de la cual puede generarse un analizador sintáctico automático, para el análisis sintáctico exhaustivo del español. En la actualidad su módulo de análisis frasal está completado[3], así como diseñado e integrado, pero hueco, su módulo de análisis clausal. El proyecto se propone rellenar los que sea posible de los submódulos de análisis clausal, recurriendo para ello a los datos extraídos de la Base de datos sintácticos BDS[4], así como, completar lo necesario para dar cuenta de todo lo que rebasa los niveles de análisis frasal y clausal: lo discursivo. En cuanto al desarrollo del módulo clausal de AVALON, quince categorías de cláusulas (secuencias de funciones organizadas en torno a una forma verbal predicado) han sido, por ahora, discriminadas, y en cada una de ellas debemos proceder aplicando el mismo método que ya se ha aplicado a una de esas categorías: las interrogativas parciales[5]. Este

[2] URL: http://www.agfl.cs.ru.nl/

[3] Descrito en María Paula Santalla, *A Formal Grammar of Spanish for Phrase Level Analysis Applied to Information Retrieval*, Lalia Series Mayor, Servizo de Publicacións e Intercambio Científico, Universidade de Santiago de Compostela, Santiago de Compostela, 2002.

[4] URL: http://www.bds.usc.es/

[5] Descrito en Iria del Río Gayo, "Sintaxis de un tipo de cláusula interrogativa a través de datos de

ISSN 1135-5948

método combina el acceso a la información incluida en gramáticas y tratados lingüísticos sobre el tipo de cláusula en cuestión con la extracción de datos organizada acerca de ese tipo de cláusulas en BDS. Esta extracción está condicionada por la estructura de AVALON y procede según la aplicación de una serie de criterios sucesivos: reducción de la información sobre argumentos en BDS a los argumentos considerados en AVALON, número de argumentos explícitos, orden de argumentos explícitos y su frecuencia, frecuencia de la combinación de presencia de clíticos, voz y esquema verbal. Las reglas formales (las alternativas de reescritura formales que la constituyen) que reescriben cada tipo de cláusulas reproducen fielmente estos datos de frecuencia a través de su disposición relativa. Un segundo aspecto a tener en cuenta aparte de en qué sentido, en cuanto a extensión, debe progresar la gramática, es el que se refiere a sobre qué bases descriptivas. Por ahora, las descripciones formales de hechos lingüísticos codificadas en AVALON se han basado, las del módulo frasal, en las descripciones (no tan atendidas por la tradición de estudios en lengua española como, por ejemplo, los grupos oracionales) de los grupos de palabras de esta clase recogidas en las gramáticas; las del módulo clausal predominantemente en datos reales extraídos de la BDS. La confluencia de ambas fuentes de información, con predilección por la derivada de datos reales, creemos que es la situación ideal en cualquier nivel de análisis, y con más razón cuanto más amplia sea la unidad sintáctica, quizá discursiva, considerada. Por ello, aparte de por su interés en sí mismo, el desarrollo de AVALON más allá del nivel de análisis clausal nos lleva al desarrollo de CSA, el otro gran objetivo de este proyecto que tratamos en el apartado a continuación.

3. Desarrollo de CSA

Entendiendo que, para el desarrollo de gramáticas formales y analizadores acordes, es necesario, además de recurrir a la tradición descriptiva de estudios lingüísticos, obtener datos reales extraídos de corpus o bases de datos lingüísticas, pretendemos, para poder acceder a datos de este tipo para las estructuras frasales y, sobre todo, para lo que se refiere a las discursivas, desarrollar un corpus sintácticamente analizado (CSA) con la misma exhaustividad (en términos de agotamiento de los elementos en ellas contenido, etiquetación y jerarquización de las unidades consideradas) de la que pretende dar cuenta AVALON: dado tal nivel de exhaustividad, el objetivo prioritario en el marco de este proyecto es, a este respecto, por encima de la producción de un determinado número de secuencias analizadas, el diseño y la creación de la estructura y los protocolos de trabajo, así como el sistema de etiquetación, necesarios para ello. Para elaborar el corpus, la idea central es utilizar el modelo de desarrollo colaborativo y controlado, si no la tecnología, propia de los wikis, analizando manualmente. De manera breve, la tecnología wiki subyace a un sitio web cuyas páginas pueden ser editadas por múltiples usuarios a través de su navegador web. Junto a ello, que es lo que lo define, el modelo de desarrollo de los wikis presenta las siguientes características que también, por razones obvias, son útiles a nuestros propósitos: por un lado, la simplicidad del entorno en el que se materializa y del lenguaje de marcado que utiliza, así como del resultado almacenado: texto plano. Por otro lado, la presencia (no en todos los wikis, pero sí en los que sustentan los proyectos más ambiciosos) de funcionalidades adicionales de control: registro histórico de usuarios y páginas, control de versiones de páginas, gestión de perfiles y grupos de acceso y gestión de ediciones concurrentes. En el proceso que concebimos poner en marcha, habrá una primera fase de selección y segmentación de cada texto que se decida incluir en el corpus, tras la cual cada una de las secuencias consideradas como susceptibles de recibir un análisis sintáctico independiente, pasará a constituir en una página de la aplicación wiki que desarrollemos un enlace a otra página que habrá de contener, cuando un usuario la edite para llevarlo a cabo, su análisis sintáctico. Una vez que se disponga de él, en una fase siguiente, este análisis será procesado, convirtiéndolo a algún formato interpretable por alguna aplicación adecuada para ello, con el fin de hacerlo visualizable de manera amigable, esto es, tratándose de análisis sintácticos, en forma de árbol.

corpus", en Isabel Moskowich, Begoña Crespo, Inés Lareo y Paula Lojo (eds.): *Language windowing through corpora / Visualización del lenguaje a través de corpus*, A Coruña, Universidade da Coruña, 2010, 703-716.