MorphOz: Una plataforma de desarrollo de analizadores sintáctico-semánticos multilingüe

Oscar García Marchena

Departamento de Lingüística.
VirtuOz S.A.
47, rue de la Chaussée d'Antin
75009París
ogarcia@virtuoz.com

Laboratorio de Lingüística Formal
Universidad Paris VII
30, Chateau de rentiers 75013 París
oscar.garciamarchena@linguist.jussieu.fr

1. Un analizador sintáctico-semántico

MorphOz es una plataforma de desarrollo de conocimientos lingüísticos que permite la confección de analizadores sintáctico-semánticos en cualquier lengua. Estos analizadores se diferencian de otros parsers en que sus análisis sintácticos están acompañados de análisis semánticos generados a partir del análisis sintáctico obtenido. Estas representaciones semánticas son independientes de la lengua, y, en principio, idénticos para frases de cualquier lengua con el mismo significado.

Las posibilidades de aplicación tecnológica de estos analizadores con capacidad de representación de significado multilingüe son variadas. Sus creadores, la sociedad VirtuOz, lo emplean para la confección de agentes de diálogo o *chatbots*: el usuario interactúa con una interfaz que transforma las intervenciones humanas en representaciones semánticas a las que puede responder proactivamente a lo largo de una conversación.

MorphOz utiliza un modelo de análisis gramatical diferente del de otros analizadores: en lugar de realizar un análisis sobre el orden lineal de la frase, genera una representación arborescente de su sintaxis profunda, abstrayendo así el orden sintagmático del análisis gramatical. Este tipo de representación parte de la gramática de dependencias (Tesnière: 1959), y está basado en un modelo lingüístico, la *Teoría Sentido-Texto* o *TST* (Mel'čuk: 1988), implementado gracias a una gramática de unificación que es también un modelo de representación lingüística reciente, la *gramática de unificación polarizada* o *GUP* (Kahane: 2004). Este sistema presenta la ventaja de ser un modelo lingüístico modular, permitiendo separar en dimensiones de análisis independientes la información morfológica, el léxico, las construcciones sintácticas, su semántica, y el orden de palabras.

2. Adaptación multilingüe

2.1. Parámetros gramaticales en tipología lingüística

Los modelos recientes en lingüística formal (HPSG, LFG, etc.) proponen una organización gramatical de la lengua al mismo tiempo, y en grados diversos, lexicalista y construccionista. La información gramatical sobre cómo se combinan las unidades de una lengua dada están codificadas en tres áreas: léxico, construcción, y orden de palabras. El léxico, identifica la (sub)categoría, el significado, y la morfología que vincula un token con un lema; las construcciones indican la estructura en la que aparece esa (sub)categoría. Finalmente, el orden de palabras señala las posibles posiciones de los argumentos.

Una vez parametradas así las lenguas, podemos formalizar el grado de gramaticalización de cada uno de estos módulos: una gramática del chino contendrá un vocabulario sin información morfológica, varias construcciones gramaticales, y pocas reglas de orden lineal, marcando así un rígido orden de palabras. Para el español, al contrario, se precisará bastante información morfológica en el léxico, y numerosas reglas de orden lineal, para formalizar la variedad de órdenes sintagmáticos posibles.

2. 2. Parámetros gramaticales en MorphOz

Siguiendo esta corriente lexico- construccionista de la lingüística formal actual, MorphOz cuenta con un sistema modular que permite separar los diferentes tipos de información lingüística,

tratarlas independientemente, e incluso transferir los parámetros comunes a otras lenguas con similitudes estructurales. De este modo, construir un motor de análisis para cualquier lengua equivale en MorphOz a distribuir adecuadamente los recursos lingüísticos en tres áreas: léxico (con indicación categorial, semántica y morfológica), construcciones, y orden de palabras.

El léxico de cada lengua es tratado como un módulo intraspasable, pero no así el inventario de categorías gramaticales; las construcciones asociadas a las categorías, y el orden de palabras son frecuentemente exportables a lenguas genética o tipológicamente cercanas.

Las construcciones gramaticales describen las dependencias sintácticas: identifica núcleos y dependientes, y las funciones gramaticales que identifican la dependencia (sujeto, OD, OI, CC, etc.). Asimismo, las construcciones contienen información semántica: a cada lexema corresponde un semema-definición, que ocupa un lugar en una ontología (basada en Wordnet), y a cada función sintáctica le corresponde un rol semántico regular (agente, tema, paciente, etc.). Si bien esta decisión es extremadamente problemática desde un punto de vista teórico, se adapta bien a los propósitos de representación semántica de la TST (Nasr: 1996).

Esta representación semántica última debe ser la misma para todas las lenguas. De este modo, la tarea final del lingüista es controlar que las representaciones semánticas de frases con significado equivalente sean idénticas en lenguas diferentes, a pesar de las diferencias en las representaciones de la sintaxis profunda (sintaxis de dependencias).

2.2.1. Construcciones

Respecto a las lenguas romances, alrededor del 80% de las construcciones han sido compartidas para la confección de gramáticas de español, italiano y portugués. Un 70% son compartidas entre estas lenguas y el francés. Las estructuras diferentes son sobre todo las (sub)categorías verbales con diferente subcategorización, a causa principalmente de la ausencia de reglas para las alternancias en la realización de valencias.

Para evitar calcos de modelos gramaticales de tradiciones lingüísticas diferentes, para otras lenguas, se integra directamente una gramática de construcciones completa, pero siempre inspirada en las soluciones ya adoptadas. Las frases averbales del chino, por ejemplo, siguen así el mismo esquema que las oraciones nominales romances, en las que el verbo copulativo no aporta significado, sino que forma un predicado con su atributo.

2.2.2. Orden de palabras

El orden de palabras está codificado siguiendo el sistema de la TST, según el cual el orden lineal corresponde a una relación de distancias a izquierda o derecha entre el núcleo y su dependiente. El paso entre la sintaxis profunda y superficial se limita a un *mapping* o proyección de las dependencias en la linealidad de la lengua. Las lenguas romances difieren sólo en algunas reglas, particularmente respecto al orden de clíticos. Otras aplicaciones conciernen las posibilidades de realización en la periferia oracional, o la pasiva en chino, que se define únicamente en función del orden de palabras.

3. Conclusión

La implementación de una teoría lingüística como la TST para la construcción de analizadores sintáctico-semánticos tiene una utilidad doble: plataforma de desarrollo para la investigación en lingüística formal, y aplicaciones industriales variadas: agentes de conversación, sistemas de comprensión multilingüe, etc.

El análisis de la sintaxis profunda proporciona además una ventaja sobre otros analizadores: al separar orden de palabras y dependencias, no corremos el riesgo de confundir complementos de adjuntos sea cual sea la posición de éstos.

4. Referencias

S. KAHANE, “Grammaires d’unification polarisées”, en *11ième Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN’04),* Fès, Maroc, France, 2004.

I. MEL’CUK, *Dependency Syntax : Theory and Practice*. Albany, N.Y., The SUNY Press, 1988.

A. NASR, *Un modèle de reformulation automatique fondé sur la Théorie Sens Texte: Application aux langues contrôlées.* Tesis Doctoral en informática, Universidad Paris 7, 1996.

L.TESNIÈRE, “Comment construire une syntaxe” en *Bulletin de la Faculté des Lettres de Strasbourg*, 1934, 7 - 12, 219–229.