PROYECTOS

El proyecto MLAP 93-015, una investigación sobre especificaciones lingüísticas para estándares industriales

Toni Badia, UPF (Barcelona)

1. Introducción

La finalidad principal de este proyecto es investigar sobre la viabilidad de reutilizar las especificaciones lingüísticas desarrolladas en el marco del proyecto Eurotra de traducción automática para aplicaciones generales de procesamiento del lenguaje natural.

En efecto, las especificaciones lingüísticas obtenidas a lo largo de los años de funcionamiento del proyecto Eurotra constituyen uno de los aspectos mejor valorados del proyecto por las comisiones independientes de valoración, especialmente, en el informe de final de proyecto. En éste, se recomienda que sean dadas a conocer y puestas a disposición de los investigadores y lingüistas computacionales en general. No obstante, estas especificaciones se encontraban en un formato demasiado idiosincrásico, puesto que estaban ligadas muy estrechamente al formalismo ET de Eurotra.

Con esta finalidad principal se ha lanzado el proyecto MLAP 93-015, con el título completo de "Investigation of linguistic specifications for future industrial standards: The Eurotra reference manual". Se trata de un proyecto colectivo en el que participan los siguientes centros de investigación:

- Institut für Angewandte Informationsforschung (IAI), Saarbrücken, Alemania
- University of Essex, Colchester, Reino Unido
- Universitat Pompeu Fabra, Barcelona
- Centre for Computational Linguistics, UMIST, Manchester, Reino Unido
- Gruppo DIMA, Turín, Italia
- Centre for Computational Linguistics, University of Leuven, Bélgica

El proyecto parte de la doble idea de que los elementos básicos de las especificaciones lingüísticas de Eurotra son válidos para una descripción multilingüe (y en gran parte también de las descripciones monolingües), y de que no obstante muchos aspectos de la formulación de estas especificaciones están justificados sólo por el formalismo idiosincrásico en que se implementaron.

Así pues, se trata de hacer disponibles los resultados de la investigación

sobre la descripción lingüística en un formato homologado por las corrientes actuales de la lingüística computacional. En los desarrollos recientes de la lingüística computacional, la notación en términos de estructuras de rasgos tipificadas se ha ido convirtiendo en un estándar para la expresión formal de los hechos lingüísticos. Por ello, el formalismo en el que se presenten los resultados estará basado en la representación en estructuras de rasgos tipificadas.

2. Objetivos generales

Los objetivos del proyecto tienen, pues, una doble orientación: presentar en un formato estándar los resultados del proyecto Eurotra, y mejorar aquellos aspectos que dependían demasiado del formalismo usado. En concreto, los objetivos del proyecto pueden resumirse en estos cinco:

a/ El primer objetivo, relativamente humilde, es el de redescribir el conocimiento lingüístico contenido en el Manual de Referencia de Eurotra en un formato más asequible. Naturalmente, esto es posible solamente en relación a los aspectos de las especificaciones que no dependen fundamentalmente de los aspectos más idiosincrásicos del formalismo (estructuras de datos y herramientas computacionales específicas). Idealmente, este objetivo se obtiene eliminando de las especificaciones las referencias y dependencias respecto del formalismo de Eurotra.

b/ El segundo objetivo, más ambicioso, consiste en rediseñar los contenidos que no pueden ser simplemente reescritos. Esto afecta principalmente a los fenómenos en que el tratamiento está ligado muy estrechamente al formalismo ET, de manera que el rediseño incluye cambios importantes de formalizaciones en ET a representaciones estándares en LC:
- estratificación vs. estructuras de rasgos multidimensionales
- análisis transformacional vs. compartir estructuras
- tratamiento por regla vs. tratamiento léxico

c/ El conjunto de fenómenos que han sido reescritos deben ser ampliados, de manera que en la medida de lo posible se puedan llenar las lagunas (en profundidad y en extensión) propias de la descripción original. Conviene tener presente que en muchos casos, estas lagunas eran consecuencia de la falta de expresividad del formalismo ET.

d/ Este proyecto proporciona información sobre las descripciones lingüísticas al proyecto LSGRAM, cuyo objetivo principal es desarrollar recursos gramaticales en un formalismo especifico, Alep (ver descripción de este proyecto en este mismo número).

e/ Puesto que las especificaciones originales de Eurotra estaban en gran parte fundamentadas desde un punto de vista multilingual, es interesante que el nuevo proyecto mantenga ente interés por la multilingualidad y por la transferencia. El proyecto, por lo tanto, pretende investigar en lo posible los requisitos que debe tener una descripción de la relación de traducción en formalismos de estructuras de rasgos tipificadas.

Después de todos los estadios de desarrollo del proyecto y de las etapas finales de armonización de los resultados obtenidos, se espera que el proyecto proveerá una descripción integrada (en la medida de lo posible) de una gran parte de los fenómenos básicos de la descripción lingüística de las lenguas tratadas en Eurotra.

El proyecto empezó en enero de 1994, con una fase preparatoria que debía establecer los mínimos necesarios para adoptar un formalismo común. Posteriormente, se entró en la etapa de la reescritura de los aspectos que podían ser más fácilmente incorporados a los nuevos formalismos. Finalmente, en estos momentos, los distintos equipos que trabajan en el proyecto están dedicando sus esfuerzos a los objetivos de rediseño y ampliación, principalmente.

Proyecto GEISA: GEstión Integrada de Sinónimos y Antónimos

Investigador Principal: O. Santana. Colaboradores: J. Pérez, S. Santos, G. Rodríguez, Z. Hernández. Grupo de Investigación en Estructuras de Datos. Departamento de Informática y Sistemas. Universidad de Las Palmas de Gran Canaria

Realización de una aplicación de gestión de sinónimos y antónimos en español que tenga en cuenta los accidentes gramaticales.

Objetivos:

a) Selección de la cabecera del diccionario y reconocimiento de los accidentes gramaticales que la relacionan con la palabra original.

b) Flexión del sinónimo o antónimo seleccionado.

c) Almacenamiento estructurado de un diccionario de sinónimos y antónimos para una óptima ocupación y tiempo de respuesta.

d) Interacción con el diccionario en un entorno amigable.

Selección de la cabecera del diccionario y reconocimiento de los

accidentes gramaticales que la relacionan con la palabra original.

Por medio de un catálogo de terminaciones se extraen de la palabra original sus posibles terminaciones. Para cada una de las raíces que resultan, se lleva a cabo su búsqueda en el diccionario (organizado por raíces) y se comprueba si admite o no la terminación que supone en la palabra original; en los casos exitosos se habrán conseguido las cabeceras del diccionario y las flexiones que las relacionan con la palabra original. Debido a que la palabra original puede ser también un derivado verbal (conjugado, sustantivado, adjetivado o adverbializado) se somete al módulo de reconocimiento de verbos que averigua si es o no; en caso afirmativo, proporciona las cabeceras y las flexiones correspondientes.

Flexión del sinónimo o antónimo seleccionado.

Cuando el usuario identifique el sinónimo-antónimo apropiado se aplican por defecto las flexiones que correspondan (en caso de que no sea posible se indica la razón). Adicionalmente, se permite variar la flexión de acuerdo con las características gramaticales de la palabra elegida.

Organización del diccionario:

El diccionario se organiza por raíces. Cada cabecera consta de un vocablo, dividido en raíz y terminación principal, y de su categoría gramatical; podrá ir acompañada de sus terminaciones básicas (para cambios de género y número), de una indicación de derivación y de una lista de terminaciones alternativas (para formación de aumentativos, diminutivos, despectivos, sustantivación, adjetivación, adverbialización o pronominalización). De esta forma se dispone del conocimiento necesario para la flexión.

Interacción con el diccionario en un entorno amigable.

En una primera fase se muestran las cabeceras que resultan de la búsqueda junto al número de sinónimos que ostentan Äla cabecera más parecida a la palabra original aparece preseleccionada. El usuario, al seleccionar la que le interesa, consigue la correspondiente lista de sinónimos-antónimos y el tamaño del conjunto de los sinónimos de éstos. Mediante las listas de sinónimos o antónimos se lleva adelante la navegación de forma muy asequible al usuario.

Flexiones consideradas:

¨ Cambios de género y de número en sustantivos y adjetivos.

¡ No cambian su morfología con el género: los sustantivos de género común, epiceno o ambiguo ni los adjetivos de una terminación.

¡ Algunos sólo se usan en plural o en singular y otros son invariables respecto al número.

¡ Desde el punto de vista de la sinonimia es preciso considerar que

algunos sustantivos cambian su significado con el género o con el número y otros sufren heteronimia.

¨ Los adverbios carecen de género y número.

¨ Aumentativos, diminutivos y despectivos en los sustantivos y adjetivos.

¨ Grado superlativo del adjetivo con sus irregularidades.

¨ Sustantivación: Con los adjetivos que terminan en *-ble* se construyen los sustantivos terminados en *-bilidad.*

¨ Adjetivación: Con algunos sustantivos femeninos se construyen adjetivos terminados en *-idad.*

¨ Adverbialización: Con la forma femenina de los adjetivos se construyen adverbios de modo al añadir la terminación *-mente.*

¨ En los verbos: formas conjugadas (persona, tiempo, modo y voz); formas pronominales, sustantivadas, adjetivadas y adverbializadas.

n No se flexionan preposiciones, conjunciones, interjecciones, locuciones o frases, ni palabras de otros idiomas.

Formación de raíces y terminaciones

Se toma como raíz de una palabra la parte de la cadena de caracteres que permanece invariable frente a las posibles operaciones de flexión; en los verbos se considera la raíz gramatical Äusan un módulo propio de reconocimiento y flexión. El resto de la palabra compone su terminación principal. No se consideran invariables la aparición o desaparición de la tilde, ni el cambio de su posición, así como las alteraciones morfológicas (*z* por *c*, *g* por *j*, *c* por *qu*, etc.). La raíz es vacía si la flexión de la palabra produce alteraciones en su primer carácter (en caso de aparición de una tilde) y la terminación principal lo es cuando la raíz alcance a toda la palabra (en caso de no admitir flexión). Téngase en cuenta que una forma flexionada se construye concatenando la raíz con la terminación de la flexión correspondiente.

La lista de terminaciones básicas proporciona las del masculino singular, femenino singular, masculino plural y femenino plural respectivamente Äse indican oportunamente las flexiones inexistentes y las vacías. Estas terminaciones se obtienen a partir de la palabra, de la información acerca de su categoría gramatical, de su número de terminaciones, de las reglas correspondientes para cambiar el número y sus excepciones. Existen además otras listas de terminaciones análogas para la formación de aumentativos, diminutivos, despectivos, superlativos, sustantivaciones, adjetivaciones y una sóla terminación para la adverbialización.

Gramática electrónica de los operadores verbales de primer nivel en español

Roser Palacios Porras
Laboratorio de Lingüística Informática
Dep. Filología Española
Universidad Autónoma de Barcelona

1. Introducción

Nuestro trabajo se integra en el proyecto de desarrollo de un Sistema de Diccionarios y Gramáticas Electrónicos del Español (SDGEE), iniciado en el Laboratorio de Lingüística Informática (en adelante, LaLI) de la Universidad Autónoma de Barcelona en 1986, bajo la dirección del Dr. Carlos Subirats. Nuestra investigación consiste en el estudio en el léxico de propiedades formales y semánticas de los operadores (cf.Harris 1982; 1991) verbales de primer nivel en español. Los resultados de nuestra investigación se incorporarán a la base de datos que se está desarrollando en el LaLI, y constituirán una subgramática que se incorporará como un nuevo módulo dentro del Sistema de Gramáticas Electrónicas del Español (SGEE). El SGEE es una base de datos en RDB con SQL, que corre en el sistema operativo VAX/VMS. La información lingüística de la base de datos del SGEE se encuentra en ficheros ASCII, y por lo tanto es independiente tanto del sistema operativo como de la base de datos que lo carga.

2. Desarrollo del proyecto

Partiendo del léxico de operadores verbales del Diccionario Electrónico de Formas Simples del Español (DEFSE), uno de los módulos constituyentes del Sistema de Diccionarios Electrónicos del Español (SDEE) (Subirats 1989, 1992, 1995), hemos empezado a construir una clase de operadores verbales (Harris 1991) integrada apróximadamente por 2.000 elementos léxicos, que van a constituir el objeto de estudio de nuestra investigación. Esta clase está formada por operadores de primer nivel, es decir, verbos que no admiten un operador entre sus argumentos; se trata, por lo tanto, de los verbos que no admiten una oración subordinada, como p. ej., *abrir*, *estrangular*, etc.

Hemos determinado las clases de dependencias, es decir, el régimen de los operadores verbales, o lo que es lo mismo, el requerimiento argumental. Así, el operador verbal *dormir* se incluye dentro de la clase de dependencia *On* porque exige un único argumento, p. ej., *Juan duerme*. Una vez determinadas las clases de dependencias, hemos establecido las estructuras de base de dichos

operadores. Así, el operador verbal *dormir* proyecta su requerimiento argumental en la estructura de base *N1 V*. La determinación de la estructura de base, es decir, la proyección lineal del operador, nos ha permitido separar operadores que pertenecen a la misma clase de dependencia. Los verbos *matar* y *carecer* son operadores de dos argumentos que pertenecen a la clase de dependencia *Onnn*; sin embargo, la estructura de base de *matar* es *N1 V N2*, p. ej., *Pedro mató un león africano*, y la estructura de base de *carecer* es *N1 V P N2*, p. ej., *Pedro carece de principios*. Así pues, hemos establecido, entre otras, las siguientes clases de operadores verbales de primer nivel:

Requerimiento argumental	*Ejemplo*	*Estructura de base*
O	**llover**	*V*
On	**dormir**	*N1 V*
Onn	**matar**	*N1 V N2*
	carecer	*N1 V P N2*
Onnn	**nombrar**	*N1 V N2 N3*
	regalar	*N1 V N2 a N3*
	tildar	*N1 V N2 de N3*

Asimismo, hemos establecido subclases de operadores en función de los marcadores de argumento, es decir, en función de las preposiciones que introducen los diferentes argumentos (por ejemplo: *absolver de*, *convertir en*). Esto nos ha permitido separar operadores como *absolver*, *convertir* o *dar*, cuyo requerimiento argumental es el mismo, es decir que pertenecen a la misma clase de dependencia (en este caso pertenecen a la clase *Onnn*), pero cuya estructura de base es diferente, constituyendo así tres subclases de operadores distintos:

Onnn	**absolver**	*N1 V a N2 de N3*
	convertir	*N1 V a N2 en N3*
	dar	*N1 V N2 a N3*

En tercer lugar, hemos determinado las propiedades de selección y transformacionales sobre las que estamos centrando nuestro estudio. Hemos aplicado el concepto de *entrada* de la gramática léxica del español (Subirats 1987) a la clase de verbos que estudiamos. Este primer análisis nos ha permitido separar operadores del mismo nivel con idéntico requerimiento argumental, en función del conjunto de propiedades formales distintas, que

estén asociadas a significados diferenciales. Esto es lo que nos ha permitido desdoblar en nuestra gramática electrónica la forma morfofonológica *abandonar* en siete operadores distintos.

Cada uno de los operadores verbales determinados se incluirá en la base de datos, con la especificación de sus propiedades formales y de selección.

El estudio de los operadores verbales de primer nivel que estamos desarrollando se inserta dentro de una nueva línea de investigación que tiene aplicaciones en el ámbito del tratamiento automático de la lengua española, y más concretamente en el análisis sintáctico automático (cf.Johnson 1987). Asímismo, nuestro proyecto está contribuyendo al desarrollo de una teoría cuyo objetivo final es el tratamiento automático de la lengua española y el reconocimiento y análisis automático de oraciones del español.

La labor realizada en la determinación de los operadores de primer nivel extraidos del SDEE nos ha permitido abordar los problemas del estudio en el léxico de las propiedades formales y semánticas de dichos operadores.

Numerosas cuestiones inherentes al proceso de determinación de los operadores han sido abordadas mediante el estudio de propiedades transformacionales y de selección de algunos operadores. Dichas propiedades son revisadas y verificadas de una forma sistemática durante el proceso de selección de los operadores verbales que estamos utilizando para la realización de nuestra investigación.

4. Referencias

HARRIS, Z. 1991. *A Theory of Language and Information. A Mathematical Approach.* Oxford: Clarendon Press.

HARRIS, Z. 1982. *A Grammar of English on Mathematical Principles.* New York: John Wiley.

JOHNSON, S.B. 1987. *An analyzer of the information content of sentences.* New York University, Computer Science Department, Tesis doctoral.

SUBIRATS RÜGGEBERG, C. 1987. *Sentential Complementation in Spanish. A lexico-grammatical study of three classes of verbs.* Amsterdam/Philadelphia: John Benjamins.