

Una plataforma para soporte a la investigación lingüística: Análisis morfológico de la lengua ugarítica

Ana García-Serrano, Jesús Contreras
ISYS-Hermenéutica
DIA-Facultad de Informática
Universidad Politécnica de Madrid
{agarcia,jcontrer}@isys.dia.fi.upm.es

Resumen

Herramienta para ayuda a la investigación

Se presenta una herramienta para ayuda a la investigación lingüística en el área de la morfología. Ha sido desarrollada para cubrir las necesidades de filólogos del grupo de investigación del laboratorio de Hermenéutica del CSIC, en su labor de estudio de la lengua ugarítica. El principal objetivo es ofrecer soporte automático para el planteamiento de hipótesis y revisión de las conjeturas durante el trabajo del filólogo en el estudio de una lengua desconocida. Debe permitir al filólogo la gestión y actualización de los diferentes tipos de informaciones que el filólogo maneja durante las labores propias de su investigación. El diseño y el desarrollo de la aplicación se han orientado al usuario final bajo la aproximación de un sistema basado en el conocimiento con interfaces amigables para todos los componentes de información así como de los procesos y métodos de su gestión.

Lengua ugarítica

La lengua estudiada proviene del reino de Ugarit cuyos restos se hallan en el territorio de la actual Siria y se datan del segundo milenio Antes de Cristo. Es uno de los primeros alfabetos fonéticos y la lengua ugarítica pertenece a la familia de las semíticas. El corpus disponible se halla en tablillas de cerámica escrito con letra cuneiforme, y contiene unas 60.000 palabras completas, el resto son palabras incompletas debido al deterioro del material cerámico. Tanto el léxico como todo el conocimiento sobre la morfología ugarítica se ha tomado del laboratorio de hermenéutica del centro de estudio del antiguo oriente del CSIC, que ha sido utilizado en el AMU un analizador morfológico desarrollado sobre una plataforma Macintosh [Cunchillos & Cervigón].

Metodología KSM

Se ha utilizado durante el diseño del modelo de la morfología ugarítica la metodología KSM [Cuenca & Molina 96] para estructurar la morfología en diferentes áreas de conocimiento. A partir de esta estructuración se identificó el formalismo específico de cada área concreta y descomponer las tareas en sub tareas utilizando diferentes paradigmas. Siguiendo los

pasos de la metodología KSM se identificaron las áreas de conocimiento (jerarquía), definiendo para cada una su parte declarativa (la base de conocimiento) y su parte funcional (tareas). Las unidades primarias en esta jerarquía contienen el conocimiento de dominio que es gestionado por el experto lingüista: añadiendo nuevos elementos, eliminando o revisando la validez de otros o simplemente modificando los existentes.

Las áreas primarias identificadas para el modelo de morfología son cinco. La primera contiene el léxico con la relación de lexemas atestiguados de la lengua. Se ha utilizado una tabla con las tareas de consulta, inserción y borrado asociadas. La segunda contiene las reglas de formación que permiten construir palabras morfológicamente correctas de la lengua a partir de secuencias válidas de morfemas y lexemas. Otra área es la que contiene la relación de todos los morfemas conocidos, con un formalismo similar al área del léxico. El área de clasificación gramatical utiliza un formalismo de jerarquía de marcos para clasificar la palabra de entrada según sus componentes, identificado su categoría y calculando el valor de los atributos gramaticales. El último área incorpora el conocimiento sobre los fenómenos lingüísticos que puede sufrir una palabra, como la asimilación de componentes, su transformación o pérdida en determinadas condiciones. También permite definir la forma de cálculo de los atributos gramaticales a partir de los componentes identificados.

Una vez diseñado el modelo, se instancia con el conocimiento del dominio de la lengua ugarítica en las bases de conocimiento de cada área primaria. La aplicación final utiliza el modelo diseñado, ofreciendo un interfaz amigable a cada base de conocimiento, haciendo transparente al usuario la codificación interna del conocimiento y permitiendo su fácil gestión.

Funcionalidades

El sistema permite analizar morfológicamente una palabra dada ofreciendo todas las soluciones posibles, según el conocimiento de la aplicación. Este conocimiento es gestionado por los lingüistas, de manera que lo pueden alterar, borrar o

incrementar. Cada modificación tiene efectos inmediatos en el resultado del siguiente análisis, de manera que la aplicación se puede utilizar para validar conjeturas, observando el comportamiento y los resultados. En cada análisis el sistema presenta una explicación que contiene los elementos de las bases de conocimiento que ha utilizado, justificando el resultado y el análisis propuesto. Además del análisis morfológico de una palabra, es posible definir parcialmente la entrada al sistema, dejando que éste genere las diferentes alternativas de solución. De esta forma pueden obtenerse todas las palabras que cumplen unas restricciones morfosintácticas (de género, número, etc..) o es posible especificar parcialmente los componentes de la palabra de entrada, al definirse solamente una parte y siendo el sistema el que genera todas las restitutiones morfológicamente válidas.

Al presentar al filólogo todas las soluciones que sean correctas morfológicamente, permite la validación de las hipótesis introducidas o efectuar posteriormente el análisis sintáctico automático exhaustivo al disponer de todas las alternativas morfológicas posibles.

Pasos de la demostración

En primer lugar se mostrará la funcionalidad básica del sistema, introduciendo una palabra de entrada y efectuando un análisis. Se mostrarán todas las soluciones de descomposición con sus posibles análisis según el estado del conocimiento incorporado. El sistema dispone de una base de datos donde almacena las soluciones inferidas anteriormente, de manera que el usuario puede decidir si recuperar un análisis hecho o hacer una nueva inferencia. Según las explicaciones que se ofrecen para la solución, es posible editar todos los elementos de las bases de conocimiento que hayan sido utilizadas a lo largo del proceso del análisis.

En este punto de la demostración se pueden introducir modificaciones para observar que efectos tienen sobre los resultados obtenidos. La parte menos atestiguada de la morfología ugarítica esta comprendida en el área de fenómenos lingüísticos, siendo interesante formular algunas hipótesis y poder verificar las consecuencias que produce. También es posible añadir nuevos elementos descriptivos que permitan clasificar palabras de entrada, añadir borrar o modificar lemas de composición o modificar lexemas o morfemas. Una modificación interesante consiste en añadir al léxico lexemas de otras lenguas similares (hebreo antiguo, fenicio, etc..) y ver como se comporta bajo las restricciones de la morfología ugarítica.

Dado el estado de las tablas, el filólogo se ve obligado a hacer restitutiones de palabras incompletas. El sistema permite definir palabras incompletas en su entrada, indicando que algunos componentes tienen forma indefinida e inferir así todas las posibles soluciones que satisfagan esta definición parcial. Para evitar que el sistema haga una generación intratable de miles de palabras se pueden añadir restricciones de tipo gramatical, si se conocen. Se puede limitar la categoría (verbo), o algún atributo (número, género, etc..) utilizando algunas reglas de tipo sintáctico atestiguadas. La complejidad del proceso de inferencia es lineal respecto al número de soluciones que se ofrecen, siendo recomendable, por tanto, introducir todas las restricciones conocidas en la entrada.

Ficha técnica

El diseño y la implementación del prototipo se realizó con la herramienta KSM operativa sobre Solaris OS. Esta plataforma básica se enriqueció con interfaces en Tcl/Tk.

La aplicación actual se ha realizado en Windows NT 4.0, utilizando Ciao Prolog [Hermenegildo 94] para el desarrollo de la parte de gestión de conocimiento así como de las tareas propias del sistema. El Ciao Prolog es un lenguaje en desarrollo del grupo CLIP del Departamento de Inteligencia Artificial de la UPM que permitirá el desarrollo del modelo distribuido de esta aplicación. Los interfaces de las bases de conocimiento así como la aplicación principal se han implementado utilizando Microsoft Visual Basic 5.0.

Referencias

- [Cuenca & Molina 96] J. Cuenca and M. Molina. KSM: An environment for Design of Structured Knowledge Models. Chapter of the book "Knowledge-Based Systems-Advanced Concepts, Techniques and Applications". De. Spyros G. Tzafestas. World Scientific Publishing Company.
- [Cunchillos & Cervigón 96] J.L. Cunchillos and R. Cervigón (1996) 'SIAMTU: El Analizador Morfológico Ugarítico en el Entorno 4D, Technical Report.
- [Cunchillos & Zamora 95] J. L. Cunchillos and J. A. Zamora (1995): Gramática Ugarítica Elemental, Ediciones Clásicas.
- [García-Serrano, Contreras, 98] A. García Serrano, J. Contreras "A Computational Platform for Ugaritic Morphological Analysis", LREC 98.
- [Hermenegildo 94] M. Hermenegildo, the CLIP group. Some Methodological Issues in the Design of CIAO - A Generic, Parallel, Concurrent Constraint System. Principles and Practice of Constraint Programming, pages 123-133, Springer-Verlag, LNCS, 874, May 1994.