Universidad&Ciencia
Pág. 121-133
Recibido: 28/04/2017 Aceptado: 04/07/2017
Publicado 30/08/2017

Vol. 6, No. 3, agosto-noviembre (2017)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

ESLATÍN3. TRADUCTOR DE TÉRMINOS BOTÁNICOS EN LATÍN AL ESPAÑOL

ESLATÍN3. TRANSLATOR OF BOTANICAL TERMS IN LATIN TO SPANISH

Autores: Daril Alemán Morales
Yuneisy Jimenez López
Institución: Universidad de Ciego de Ávila Máximo Gómez Báez
Correo electrónico: daleman@unica.cu

RESUMEN

Los profesionales de la botánica requieren conocer un sistema preciso y universal que les permita describir las nuevas unidades que van siendo incorporadas al sistema de su ciencia. En la Universidad Central Marta Abreu de las Villas (UCLV) existe un Jardín Botánico en el cual se llevan a cabo investigaciones acerca del mundo vegetal. Por solicitud de esta entidad surgió la idea de desarrollar una aplicación computacional que dotara a sus especialistas de la herramienta básica en la práctica taxonómica. Como antecedente de este trabajo, se desarrolló una primera versión del software con el fin de automatizar el proceso taxonómico. El mismo fue desarrollado como trabajo de Diploma por parte de dos estudiantes de la carrera Ciencia de la Computación en el año 2006.
La aplicación se desarrolló en dos etapas. En la primera de ellas se realizó el diseño del programa que daría solución al problema planteado. Para ello se utilizó la notación del Lenguaje de Modelación Unificado. En la segunda etapa se procedió a la implementación del programa utilizando conjuntamente la programación a través de un lenguaje declarativo, como lo es Prolog y la programación orientada a objeto, en este caso Java. De esta forma quedó desarrollado EsLatín3, Traductor de Español a Latín para la descripción de las especies botánicas, que constituye una herramienta muy útil y eficiente.

Universidad&Ciencia
Pág. 121-133
Recibido: 28/04/2017 Aceptado: 04/07/2017
Publicado 30/08/2017

Vol. 6, No. 3, agosto-noviembre (2017)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

Palabras clave: Traductores Automáticos, Programación Lógica, Gramáticas, Lenguaje Natural, Latín.

ABSTRACT

The professionals of the botany need to know a precise and universal system that allows them to describe the new units that are incorporated into the system of their science. In the Central University Marta Abreu of Las Villas (UCLV) exists a Botanical Garden in which researches are carried out on the vegetable world. For request of this entity, arose the idea of creating an application that will provide its specialists with the basic tool in the taxonomic practice. As precedent of this work, the first version of the software was developed in order to automate the taxonomic process. It was developed like a Diploma Paper on the part of two students of the major of Computer Science in 2006. The application was developed in two stages. The first one was the creation of the design of the program that would answer the problem. For this, there was used the notation of the Unified Modelling Language. The second stage was the implementation of the program using jointly the programming through a declarative language, as it is Prolog, and the programming faced to the object, in this case Java. This way was developed EsLatín3, a Translator from Spanish to Latin for the description of the botanical species, which constitutes a very useful and efficient tool.

Keywords: Automatic Translators, Logic Programming, Grammars, Natural Language, Latin.

INTRODUCCIÓN

Haciendo uso de la computadora como herramienta de ayuda se han logrado desarrollar aplicaciones para el procesamiento del lenguaje natural, tales como correctores ortográficos, procesadores de textos especializados, reconocedores de voz, analizadores sintácticos y traductores. Los profesionales de la botánica requieren el conocimiento de un sistema preciso y universal de nomenclatura que les permita describir las nuevas unidades o

grupos taxonómicos que van siendo incorporados al corpus que conforma el sistema de su ciencia.

Ese sistema existe y constituye el fruto del esfuerzo que desde hace siglos han desarrollado los más notables botánicos por conseguir tanto una explicación científica del mundo vegetal, una organización de su conocimiento, como una comunicación uniforme, un tecnolecto universal, que derribe barreras geográficas y diacrónicas y permita el global entendimiento acerca de este ámbito que efectivamente es patrimonio de toda la humanidad. Como ha sucedido con la mayoría de las ciencias, el sistema lingüístico elegido para conformar este tecnolecto es el latín.

En la Universidad Central Marta Abreu de las Villas (UCLV) existe un Jardín Botánico, unidad docente de la Facultad de Ciencias Agropecuarias, en el cual se llevan a cabo investigaciones acerca del mundo vegetal. Esta institución solicitó al Departamento de Letras de la Facultad de Humanidades la impartición de un curso de postgrado de Latín, que dotara a sus especialistas de la herramienta básica en la práctica taxonómica.

El proceso de registrar las nuevas especies, particularmente en la Botánica es un proceso bastante exquisito, en el cual juega un papel importante la nomenclatura botánica y la latinización de los nombres propios. Todos los nombres científicos dados a las plantas o grupos de plantas (taxones) son nombres extraídos del latín. Hasta el momento, los científicos que se ocupan de la investigación en el mundo vegetal, necesitan tener un conocimiento del latín, lo que conlleva a que su trabajo se vea sistemáticamente interrumpido a causa de ello. Realizar el proceso de nomenclatura utilizando el latín resulta, por lo demás una tarea un tanto engorrosa, en la cual el gasto de tiempo es un agravante que atenta directamente contra los resultados de la investigación. Dicha tarea se realiza manualmente, o bien con el auxilio de diccionarios, o de lo contrario se necesita de la supervisión de personal calificado, preparado profesionalmente.

Conjuntamente se puede afirmar que la eficiencia del trabajo se ve afectada en los lugares donde se realiza este tipo de estudios. En el caso del Jardín Botánico que radica en la UCLV, el personal encargado de llevar a cabo la

práctica taxonómica necesita de la ayuda de un software informático para realizar dicha tarea. Resulta de mucha comodidad que se automatice la latinización de nombres propios a través de una aplicación computacional, de la cual se derive un ahorro de tiempo, además de la liberación de los científicos respecto a la necesidad de tener un conocimiento avanzado del latín. En una primera versión se realizó una aplicación sencilla con un número reducido de funcionalidades, que sirve para realizar en un inicio algunas traducciones. En el presente software se aumenta el nivel de profesionalidad del anterior y así es posible ofrecer una herramienta capaz de realizar las labores anteriormente mencionadas.

MATERIALES Y MÉTODOS

Conceptos asociados al tema abordado:

No existe en la bibliografía una definición formal de lo que significa traducción automática (TA). Según (Hernández, 2002), en sentido estricto es "el proceso por el cual una máquina traduce un texto de una lengua a otra, subdividiendo la sintaxis, identificando las partes del discurso, intentando resolver eventuales ambigüedades y, por último, traduciendo los componentes y la estructura en la lengua de destino".

Existen dos tipos de enfoques de traducción automática: los basados en normas y los basados en corpus. Las estrategias basadas en normas, de acuerdo con (Hutchins, 2007), se pueden dividir en tres enfoques tradicionales, a saber: el sistema de traducción directa, el sistema interlingua y el sistema de transferencia. El sistema de traducción directa es el enfoque más sencillo (Craciunescu and Gerding-Salas, 2007). Está diseñado para un par de lenguas determinadas. Se traduce directamente de la lengua fuente (LF) a la lengua meta (LM), su supuesto básico es que el vocabulario y la sintaxis de los textos de la lengua fuente no necesitan ser analizados, sólo lo estrictamente necesario para la resolución de ambigüedades. Normalmente, estos sistemas consisten en un único diccionario bilingüe y un programa único para analizar el texto fuente.

El segundo sistema básico es el sistema interlingua, que asume que es posible convertir un texto de LF en representaciones sintácticas y semánticas comunes para más de una lengua (Zapata and Benítez, 2009). El texto en la LF se transforma en un lenguaje intermedio mediante el componente «Análisis"». El texto en la LM se obtiene a partir de la representación del texto en el lenguaje intermedio, mediante el componente «Generación"». La estructura del lenguaje intermedio llamado «interlingua», es independiente de la lengua fuente y de la lengua meta y está basada en una lengua artificial, como por ejemplo el esperanto. Un argumento a su favor es el efecto de economía: con un solo sistema se puede traducir a varias lenguas, aunque por otro lado, la construcción de tal interlingua es un trabajo muy complejo.

El sistema de transferencia establece una representación intermedia entre las lenguas origen y meta, alrededor de la cual se organiza el análisis y la síntesis. La transferencia separa el proceso de traducción en tres fases: análisis, transferencia y síntesis y, a su vez, se puede producir en varios niveles: léxico, sintáctico y semántico (Amores, 2002).

Los enfoques basados en corpus, también llamados «enfoques empíricos», se pueden distinguir en dos sistemas, a saber, la TA basada en ejemplos (Diéguez, 1998) y la TA estadística (Somers, 2003). La idea básica de la traducción automática basada en ejemplos es reutilizar muestras reales con sus respectivas traducciones como base de una nueva traducción (Muñoz and Ramírez, 2010). Se caracteriza por encontrar traducciones correspondientes a una base de datos de traducciones reales. El proceso consiste en tres etapas: encontrar correspondencias, alinear y recombinar. En la primera etapa el sistema encontrará mediante correspondencias con la entrada muestras de traducciones que pueden contribuir a la traducción. La segunda etapa, la alineación, consiste en identificar las partes útiles de la traducción correspondiente. El tercer paso, llamado recombinación, recombina las partes correspondientes. Cuando se sabe qué partes de los ejemplos se reutilizan, es preciso intentar que las partes correspondan de manera legítima.

En su forma auténtica, la traducción automática estadística no usa datos lingüísticos tradicionales. La esencia de este método es alinear frases, grupos

de palabras y palabras individuales de textos paralelos y calcular las probabilidades de que una palabra en una frase de una lengua se corresponda con una palabra en una frase de una traducción con la que está alineada. Dado que la TA estadística genera sus traducciones a partir de métodos estadísticos basados en corpus de textos bilingües, la disponibilidad de un corpus grande de traducciones fiables es una característica esencial de este sistema. Se suele ver este método como anti-lingüístico. La idea de este sistema es modelar el proceso de traducción en términos de probabilidades estadísticas (Gironés, 2003).

Los sistemas de traducción han sido criticados por sus grandes limitaciones frente a un traductor humano, sin embargo los programas actuales pueden producir traducciones aproximadas que ayudan a los seres humanos a determinar la relevancia de un texto en un idioma extranjero determinado (y su traducción por parte de un ser humano) o pueden servir de esbozo de traducción para un editor humano. En algunas áreas de conocimiento bastante limitadas como los servicios de pronóstico del clima, los softwares de traducción automática pueden generar versiones bastante adecuadas de un texto producido inicialmente en otro idioma.

Un paradigma de programación representa un enfoque particular o filosofía para la construcción del software. No es mejor uno que otro, sino que cada uno tiene ventajas y desventajas. También hay situaciones donde un paradigma resulta más apropiado que otro. Para el caso del tratamiento del lenguaje natural el paradigma más apropiado es el lógico (Warren, 1983, Bratko, 1986). Este paradigma que resultó una apasionante novedad en la década del 70, tiene como característica diferenciadora el hecho de manejarse de manera declarativa y con la aplicación de las reglas de la lógica.

En la realización de traductores automáticos, el lenguaje de programación lógica Prolog resulta de mucha utilidad, debido a que ofrece facilidades para representar y utilizar el conocimiento que se tiene sobre un determinado dominio. Prolog es un lenguaje de programación específico para el procesamiento del lenguaje natural (Rowe, 1988). El método de funcionamiento utilizado por Prolog es un método de razonamiento deductivo muy similar al

Pág. 121-133 ISSN: 2227-2690 RNPS: 2450
Recibido: 28/04/2017 Aceptado: 04/07/2017 http://revistas.unica.cu/uciencia
Publicado 30/08/2017

razonamiento humano, aplicado sobre el conjunto de fórmulas lógicas que componen el programa. En la base de Prolog se encuentra la idea de que un algoritmo puede dividirse en dos partes: la determinación de QUÉ hay que resolver (lógica del programa) y la indicación de CÓMO utilizar el conocimiento contenido en el programa para obtener la solución buscada (control del programa). El programador únicamente ha de preocuparse de la lógica del programa, dejando que el lenguaje se ocupe del control. El control es siempre el mismo y viene dado de forma automática por la implementación del lenguaje.
La representación morfológica del texto es en este caso, un conjunto de hechos (siempre verdaderos) en Prolog, que da los lemas y las propiedades de cada palabra del texto. Esto significa que a cada palabra en español corresponde un hecho en el programa donde aparece la palabra traducida en latín y sus propiedades. En este paso, se construye un diccionario morfológico.
El latín, constituye una «lengua muerta», es decir, no existen hablantes de la lengua propiamente dicha y que la creación natural de su corpus se detuvo, haciendo de esta lengua un registro cerrado. La nomenclatura de la taxonomía botánica ha seguido la suerte del tecnolecto de la mayoría de las ciencias, teniendo como base las lenguas griega y latina. En el caso de la Botánica, el griego aparece a través de una modificación de su sistema lingüístico al sistema latino.
Una información general acerca del sistema de la lengua latina, con una adecuada orientación metodológica, podrían colocar al especialista de la Botánica en condiciones de realizar taxonomías de alta calidad lingüística con la ayuda de las herramientas que ofrecen los resultados de este trabajo.

Las gramáticas en Prolog

Un lenguaje puede verse como un conjunto (normalmente infinito) de frases de longitud finita. Cada frase está compuesta de símbolos de algún alfabeto, según una combinación determinada para formar frases correctas. Para especificar cómo construir frases correctas, en cualquier lenguaje se utiliza como formalismo las gramáticas, que constituyen un conjunto de reglas que definen la estructura legal en un lenguaje.

Las reglas de una gramática definen qué cadenas de palabras o símbolos son oraciones válidas de la lengua. Además, la gramática generalmente brinda algún tipo de análisis de la oración, en una estructura que hace su significado más explícito.

En las Gramáticas de Cláusulas Definidas (GCD) (Sterling, 1994) se permite a los no terminales contener argumentos que representen la interdependencia de los componentes de una frase.

Lenguaje de programación Java:

Para desarrollar la aplicación se hizo un análisis previo con el objetivo de determinar cuáles eran las herramientas computacionales más eficaces que permitieran una solución óptima, además de llevar a la práctica los conocimientos adquiridos durante la carrera. El lenguaje Java fue escogido, en parte, por poseer un poderoso conjunto de bibliotecas gráficas, las cuales facilitan el trabajo en la creación de la interfaz del usuario. Además, es un lenguaje sumamente fácil de aprender, permite crear sistemas más complejos y posee características como la posibilidad de reutilizar el código en diferentes partes del programa.

Métodos teóricos:

Analítico-Sintético: Se utiliza para analizar teorías y elementos bibliográficos relacionados con los traductores automáticos, permitiendo la extracción de los elementos más importantes que dan inicio a la investigación.

Modelación: Es utilizado para representar gráficamente la solución que se propone.

Métodos empíricos:

Observación: Se emplea para estudiar las características y comportamientos de las soluciones similares, permitiendo

obtener información relevante sobre el proceso de traducción automática.

Universidad&Ciencia
Pág. 121-133
Recibido: 28/04/2017 Aceptado: 04/07/2017
Publicado 30/08/2017

Vol. 6, No. 3, agosto-noviembre (2017)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

Características de EsLatín3

Este software realiza traducciones de oraciones del idioma español al latín. También realiza actualizaciones al diccionario que sirve como herramienta a la aplicación. La interfaz con el usuario resulta amigable y a la vez sencilla al ejecutar cualquier operación, además de permitirle al usuario una rápida familiarización con la herramienta (ver Figura 1). Se incluye un módulo visual de utilidad para reconocer rasgos propios de las plantas, y además sirve para orientar mejor cualquier tipo de búsqueda.

Cuando se llevó a cabo la implementación del sistema se tomaron en cuenta cuáles eran las herramientas computacionales más eficaces para lograr una solución que fuera óptima para el usuario. Como resultado de este trabajo se obtuvieron dos componentes principales: una aplicación ejecutable implementada en Java (EsLatín3.jar), un archivo «Diccionario.txt» y otro llamado «gramática.pl» implementado en SWI-Prolog. Además se utilizó la biblioteca de enlace «jpl» para lograr el enlace entre estos módulos.

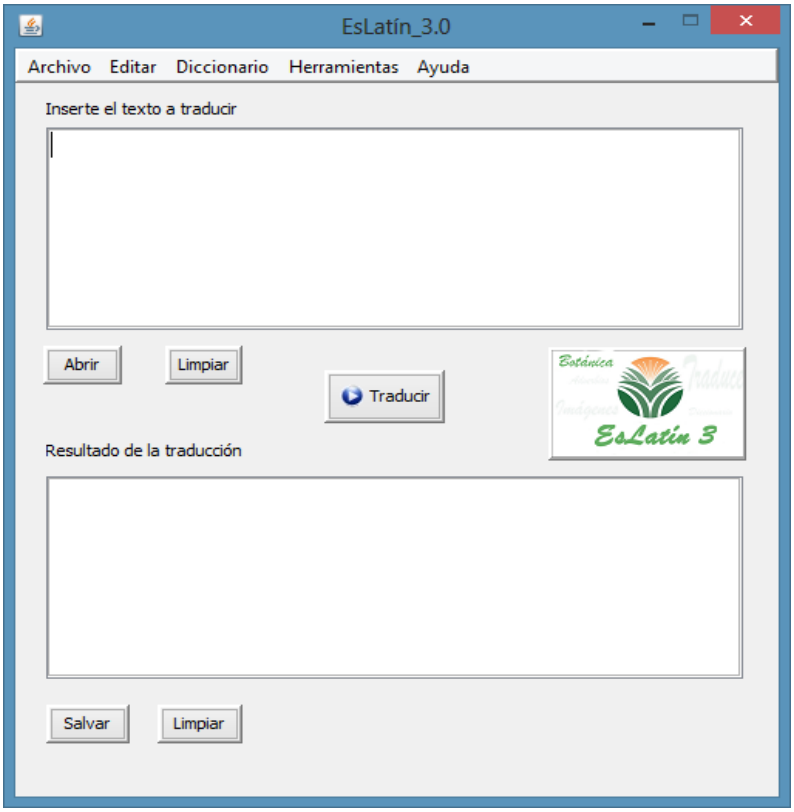

Fig. 1. Vista principal de la aplicación.

Pág. 121-133 ISSN: 2227-2690 RNPS: 2450
Recibido: 28/04/2017 Aceptado: 04/07/2017 http://revistas.unica.cu/uciencia
Publicado 30/08/2017

La implementación de EsLatín3.jar se llevó a cabo en NetBeans IDE 7.1, que es una herramienta de programación muy nutrida de componentes, las cuales resultan de mucha utilidad al realizar aplicaciones visuales y que permiten lograr ambientes agradables a la vista del usuario. El archivo «gramática.pl» contiene la gramática que es la encargada de realizar el análisis sintáctico del texto en español y además contiene predicados que son necesarios para realizar dicho análisis. En «Diccionario.txt» se encuentran todas las palabras con que cuenta el programa y sus correspondientes significados en latín.

La gramática que se implementó es una Gramática de Cláusulas Definidas (GCD), pues permite establecer concordancias de género y número entre los diferentes constituyentes de la oración. El símbolo distinguido de la gramática utilizada en la aplicación es oracion.

oracion --> cat1(Gen,Num),cat4(_,_),cat3,cat2(Gen,Num).

Nótese que se utiliza el símbolo --> para separar la cabeza del cuerpo de la regla. Esta es una característica de las GCD.

Es importante señalar que el símbolo distinguido puede tener más de un cuerpo, de hecho, la lógica lo indica así, mientras más cuerpos se tengan, más oraciones se pueden analizar con la gramática.

oracion --> sust (Gen, Num).

oracion --> adj (Gen, Num).

oracion --> adv.

Luego se definen los símbolos no terminales, que en este caso son las categorías de la oración así como los sustantivos, adjetivos, adverbios.

cat1 (Gen, Num) --> art (Gen, Num), sust (Gen, Num).

cat1(Gen,Num)--> sust(_,Num),prep, art(Gen,Num),sust(Gen,Num).

cat2 (Gen, Num) --> adj (Gen, Num), cat2a (Gen, Num).

cat3 --> adv.

cat4 (Gen, Num) --> prep, sust (_, _).

cat4 (Gen, Num) --> prep, art (Gen, Num), sust (Gen, Num).

Sust (m, s) --> [pétalo]; [ápice]; [invierno]; [ángulo].

Adj (f, p) --> [estrechas]; [ramosas]; [surcadas].

adv --> [pálidamente]; [siempre]; [densamente].

Universidad&Ciencia
Pág. 121-133
Recibido: 28/04/2017 Aceptado: 04/07/2017
Publicado 30/08/2017

Vol. 6, No. 3, agosto-noviembre (2017)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

Art (m, s) --> [el].

En el caso:

cat1 (Gen, Num) --> art (Gen, Num), sust (Gen, Num).

indica que la categoría 1 tiene género Gen (que sería m para masculino y f para femenino) y número Num. (s para singular y p para plural) si está compuesto por art(Gen,Num) y sust(Gen, Num), es decir, por un artículo y un sustantivo cuyos género y número no sólo deben coincidir entre sí, sino que serán el género y el número que se transfieran a la categoría.

Por último, comentar sobre el uso que se ha hecho del punto y coma que en Prolog tiene un valor disyuntivo. En realidad, una regla como:

Adj (f, p) --> [estrechas]; [ramosas].

equivale a:

Adj (f, p) --> [estrechas].

Adj (f, p) --> [ramosas].

Con lo que podrían haberse incluido las dos últimas en vez de la primera y la gramática sería equivalente (en el sentido de que generaría y reconocería las mismas oraciones). Sin embargo, para que sean más cortos los ficheros se ha preferido emplear el punto y coma.

RESULTADOS Y DISCUSIÓN

Como primer resultado de esta investigación se realizó un estudio del estado del arte de los traductores automáticos, permitiendo conocer las herramientas, lenguajes de programación y requerimientos de hardware de las soluciones informáticas existentes en el mundo. Debido a la utilización del lenguaje Prolog para definir la gramática, el tiempo de desarrollo de la solución se redujo considerablemente.

CONCLUSIONES

La amplia experiencia en la enseñanza del latín por el Departamento de Letras de la UCLV, permitió implementar un software que facilita el proceso de enseñanza-aprendizaje del latín para aquellas personas que constantemente hacen uso de él y que no son expertas en la materia. La aplicación quedó

provista de un diccionario con las traducciones de las palabras con las que trabajan los botánicos en la nomenclatura, tales como sustantivos, adjetivos y adverbios. El mismo puede ser modificado con facilidad por una persona que sea experta en latín.

BIBLIOGRAFÍA CONSULTADA

AMORES CARREDANO, G.: «Un mecanismo de transferencia para LFG DCG-Prolog», *Revista Procesamiento del Lenguaje Natural*, 2002.

BRATKO, I.: *Prolog Programming for Artificial Intelligence*, Ed. Addison-Wesley, 1986.

CRACIUNESCU, O. Y GERDING-SALAS, C.: *Traducción automática y asistida: ¿nuevas formas de traducir?*, S. Stringer-O'Keeffe: pp. 17, 2007.

DIÉGUEZ M.; RIEDEMANN, M.I. Y KARIN, H.: *Análisis del error en la traducción automática: algunos ejemplos de las formas -ing del inglés al español*, pp. 19, 1998.

TOMÁS GIRONÉS, J.: *Traducción automática de textos entre lenguas similares utilizando métodos estadísticos*, Tesis Doctoral Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, Valencia, 2003.

HERNÁNDEZ, P.: *En torno a la traducción automática*, Ed. Cervantes, España, 2002.

HUTCHINS, J.: *Machine translation: a concise history*. Disponible en http://www.hutchinsweb.me.uk/CUHK-2006.pdf. Visitado el 12 de febrero de 2013.

MUÑOZ, J.M.M. Y RAMÍREZ, M.V.: «Aplicaciones de traducción basadas en memorias de datos: desarrollo y perspectivas de futuro», *Entreculturas: revista de traducción y comunicación intercultural*, pp. 109-123, 2010.

ROWE, N. C.: *Artificial Intelligence Through Prolog*, Prentice-Hall, 1988.

SOMERS, H.: *Machine translation: latest developments*, Oxford handbook of computational linguistics., Oxford, Oxford University Press, 2003.

STERLING, S.: *The art of Prolog*, MIT Press,1994.

Vol. 6, No. 3, agosto-noviembre (2017)
ISSN: 2227-2690 RNPS: 2450
http://revistas.unica.cu/uciencia

WARREN, D.H.D.: *An abstract Prolog instruction set. California*, SRI International, 1983.

ZAPATA, C. Y BENÍTEZ, S.: «Interlingua: a-state-of-the-art overview», *Revista Facultad de Ingeniería Universidad de Antioquia*, (47), pp. 117-128, 2009.