Rev. Fac. Ing. Univ. Antioquia N.º 47 pp. 117-128. Marzo, 2009

Interlingua: análisis crítico de la literatura

Interlingua: a-state-of-the-art overview

Carlos Zapata[*]*, Servio Benítez*

Grupo de Lenguajes Computacionales. Escuela de Sistemas. Facultad de Minas. Universidad Nacional de Colombia, sede Medellín. Carrera 80 N.º 65-223 Of. M8A-310, Medellín, Colombia

(Recibido el 25 de marzo de 2008. Aceptado el 6 de noviembre de 2008)

Resumen

Una interlingua, es cualquier lenguaje artificial o seminatural que tenga como principales características precisión, neutralidad, inambigüedad y algún grado de formalidad para expresar las ideas que se deseen comunicar. Estas características, la convierten en una herramienta útil para resolver problemas en áreas como la traducción automática, el procesamiento del lenguaje natural y la inteligencia artificial. En este artículo, se realiza un análisis crítico de las interlinguas, incluyendo los principales conceptos al respecto, sus aplicaciones y los proyectos realizados, de tal forma que pueda servir como punto de partida para el desarrollo de nuevos proyectos en el tema.

----- *Palabras clave:* Interlingua, traducción automática, representación del conocimiento, procesamiento de lenguaje natural, inteligencia artificial.

Abstract

An Interlingua is any artificial or semi-natural language, with main features like precision, neutrality, non-ambiguity and some kind of formalism to express communicative ideas. These features have converted interlinguas in useful tools for solving problems in areas like machine translation, natural language processing, and artificial intelligence. In this paper, we make an Interlingua overview, which include concepts, applications, and developed projects about it. We hope this will be the starting point of new project development around this issue.

----- *Keywords:* Interlingua, machine translation, knowledge representation, natural language processing, artificial intelligence.

[*] Autor de correspondencia: teléfono: + 57 + 4 + 425 53 74, fax: + 57 + 4 + 425 52 27, correo electrónico: cmzapata@unalmed.edu.co. (C. Zapata)

Introducción

Una interlingua es, en su definición más sencilla, cualquier lenguaje artificial (como el lenguaje matemático) o seminatural (como el esperanto) que cuente, entre sus principales características, con precisión, neutralidad, inambigüedad y cierta formalidad para expresar las ideas que se deseen comunicar. Esto permite, como su nombre lo indica, que se convierta en una lengua intermedia a través de la cual se puede entablar la comunicación entre varios lenguajes [1- 3]. El esperanto, por ejemplo, es un reflejo de este tipo de lenguaje, pues se creó con el objetivo de que los hablantes de idiomas como el inglés, español, francés y alemán (y en general, todos los idiomas con raíces latinas o germánicas) se pudieran entender entre ellos sin necesidad de aprender muchos idiomas y con muy poco esfuerzo de aprendizaje [4]. Además, cualquier lenguaje natural (por ejemplo el inglés) se podría usar como interlingua, aunque esa no es, por definición, su función principal. Incluso, se podría suponer que algunos ideogramas basados en sistemas de símbolos (como el lenguaje chino, por ejemplo) y que conectan varios lenguajes separados pueden actuar como especies de interlinguas. Este artículo se centra, sin embargo, en las interlinguas declaradas como tal y su uso en diferentes áreas.

Las interlinguas, se usan dentro del área de procesamiento de lenguaje natural (PLN) en ramas como la traducción automática (TA), la representación del conocimiento (RC), la desambiguación de palabras, la resolución de anáforas y la generación de respuestas a preguntas formuladas por el usuario. Entre los proyectos realizados en TA, se cuentan: *UNL* [5], *ATLAS-II* [6], *PÍVOT* [7], *ROSETTA* [8] y *DLT* [8], entre otros. De hecho, en el área de la TA existe una rama llamada *sistemas basados en interlingua* para denotar aquellos sistemas computacionales que utilicen un *lenguaje intermedio* antes de realizar la traducción final. Además, este método de TA ha evolucionado a los llamados *sistemas basados en conocimientos* [9]. Algunos de estos sistemas son: *KANT* [10] y *MIKROKOSMOS* [11]. Los sistemas basados en interlingua, tienen ventajas y desventajas. Entre sus ventajas, están: la reducción del número de diccionarios necesarios para traducir entre varias lenguas, la facilidad para incluir nuevas lenguas en el sistema, la separación del conocimiento entre la lengua de origen y la lengua de destino, una mejor adecuación al proceso de traducción y la eliminación de gramáticas de comparación. Entre sus desventajas más significativas se cuentan: las dificultades para la definición de una interlingua y la necesidad de construcción de una ontología que la soporte [2, 5, 12].

En el presente artículo se describen las características esenciales de las interlinguas; se presentan los principales proyectos que emplean interlinguas; se discuten los puntos fuertes y débiles de los proyectos que utilizan interlinguas. Finalmente, se presentan las conclusiones y el trabajo futuro en esta área.

Definición y caracterización de una interlingua

Cada una de las lenguas del mundo, es un conjunto ordenado y sistemático de signos, símbolos y códigos que se relacionan entre sí para que los hablantes puedan expresar ideas, conocimientos, afirmaciones, proposiciones, etc. [13]. Sin embargo, cada lengua se impregna con la cultura y la forma cómo sus hablantes interiorizan e interactúan con el mundo, lo cual las matiza con características propias. Estos matices, son los que distancian a unas lenguas de otra, haciendo más difícil encontrar una representación interlingual [9]. Por el contrario, una lengua que se autodenomine *interlingua*, debe eliminar cada uno de estos matices y codificar los mensajes emitidos en un grado de abstracción mayor, donde cada una de las lenguas involucradas pueda equivaler a otra [9, 12]. Una interlingua es, por tanto, un lenguaje artificial diseñado con el propósito de representar el conocimiento de múltiples lenguas. En otras palabras, es una lengua intermedia, un puente donde los diferentes tipos de lenguaje natural (LN) pueden converger [2, 14]. Este concepto, orientó el diseño de interlinguas a tres diferentes tipos: las interlinguas basadas en lenguajes lógicos y artificiales, aquellas basadas en lenguajes seminaturales como el esperanto y, finalmente,

aquellas que proponen un conjunto de primitivas semánticas comunes a todas las lenguas convirtiéndose así en palabras universales (Universal Words, UW) [1, 3].

Una interlingua posee todos los rasgos característicos de las lenguas, tales como léxico, gramática y sintaxis, entre otras, los cuales se deben definir al momento de diseñar la interlingua. Además, se debe lograr que ésta sea tan expresiva como cualquier LN mientras cumple con las condiciones de precisión, carencia de ambigüedad, neutralidad y al menos, carácter semiformal. Además, las interlinguas deben permitir la RC, haciendo posible el estudio de las estructuras profundas del lenguaje. Por eso, aunque una interlingua es un lenguaje, no toda lengua puede ser una interlingua [2, 14].

El concepto de interlingua no es nuevo. La teoría de un lenguaje universal, tiene sus orígenes en el siglo XVII, cuando Descartes y Leibniz elaboraron teorías para la creación de diccionarios basados en códigos numéricos universales. Luego, matemáticos como Beck, Becher y Kircher trabajaron en el desarrollo de una "lengua universal" que no presentara ambigüedades y se basara en principios lógicos y símbolos icónicos. En esta misma época, apareció el trabajo de John Wilkins quien elaboró una interlingua en 1668 [8]. Estas teorías, fueron los pilares de lo que hoy se conoce como *teoría de los universales lingüísticos*, la cual sostiene que las lenguas poseen características o propiedades comunes denominadas *universales lingüísticos*. Es diferente un universal lingüístico de uno absoluto. El primero, se refiere a propiedades o rasgos que aparecen en las lenguas que se estén analizando, mientras que el segundo debe aparecer en todas las lenguas existentes. Algunos de estos universales absolutos son la negación, las reglas fonéticas, los sonidos vocálicos, las reglas gramaticales, etc. [15, 16].

Aplicaciones de las interlinguas

Las características especiales que exhiben las interlinguas, las hacen especialmente útiles en varias áreas del PLN. Sus usos se han extendido a la TA, la RC, la desambiguación de palabras, la resolución de anáforas, la generación de respuestas a preguntas formuladas en LN y la inteligencia artificial. A continuación, se presentarán algunos de los proyectos que emplean interlinguas.

Traducción automática

La TA empleando interlinguas utiliza dos enfoques: los sistemas basados en interlingua y los sistemas basados en conocimientos. Los sistemas basados en interlingua, realizan el proceso de traducción en dos fases: análisis y generación. En la fase de análisis, solamente se utiliza la lengua de origen para determinar el significado del texto y llevarlo a una interlingua. Mientras tanto, en la fase de generación, se utiliza el significado que se obtiene en la interlingua para generar el texto de destino en cualquier lengua [1, 17, 18].

Por otra parte, los sistemas basados en conocimientos son una evolución de los sistemas basados en interlingua. Estos sistemas, igualmente, realizan la traducción en las mismas fases de análisis y generación. Sin embargo, la gran diferencia consiste en que no sólo utilizan la información que se desprende del texto, sino que tienen acceso a un gran depósito de información llamado base de conocimiento [19], que contiene información del mundo o de un modelo del mundo llamado *ontología* [1, 20]. Existen también recursos que no se consideran interlinguas como tal, pero que contienen índices de palabras interconectadas actuando como ontologías; uno de estos recursos es el Eurowordnet. Algunos de los proyectos de TA que emplean interlinguas se describen seguidamente.

ATLAS II

Inicialmente, este sistema sólo traducía inglés y japonés, pero luego se añadieron módulos para traducir de japonés a coreano, francés y alemán. ATLAS II incluía un gran diccionario (586000 palabras en inglés y japonés) y, durante más de 10 años, la comunidad científica lo reconoció como el mejor sistema de TA [6, 21]. En la etapa de análisis, utilizaba un diccionario, unas reglas de análisis (gramática y semántica) y un modelo

del mundo. El primer análisis, era morfológico y consistía en dividir las oraciones en morfemas, que luego se incorporaban en una lista de análisis de nodos, donde se le agregaba información del diccionario. Después, venían los análisis sintácticos y semánticos que consistían en tomar la lista de nodos y aplicarles una serie de reglas gramaticales, las condiciones en que se debían aplicar, la prioridad de cada regla (en caso de que más de una regla se pudiera aplicar), el tipo de reglas, la adición de atributos gramaticales y las relaciones entre los nodos. El resultado de esta etapa, era una estructura conceptual, es decir, su interlingua. La interlingua de *ATLAS II*, constaba de una serie de relaciones binarias entre los conceptos (nodos) y las características relacionadas a cada uno de ellos. Esta estructura, se parecía mucho a una red semántica. Arcos y nodos constituían la red, donde los nodos eran los conceptos y los arcos las relaciones. Por "concepto" se entendían las palabras que representaban un significado (verbos, sustantivos, adjetivos, por ejemplo), mientras que las relaciones describían nexos profundos entre las palabras (como quién era el agente o el objeto en la oración). Además, existían los arcos unarios, los cuales se utilizaban para indicar características de los conceptos (tales como el tiempo del verbo). El sistema, validaba las estructuras conceptuales confrontándolas con un modelo del mundo. Este modelo, contenía todas las relaciones posibles entre los conceptos y el sistema sólo aceptaba una estructura conceptual si ésta se incluía en su modelo del mundo. De lo contrario, se hacía otro análisis de la oración. La etapa de generación, se dividía en dos procesos: el de transferencia y el de generación. El proceso de transferencia trataba de acercar la estructura conceptual obtenida a la lengua meta. Este acercamiento, consistía en escoger la traducción más natural posible, es decir, en evitar la traducción literal, y se lograba aplicando unas reglas de transferencia a la estructura conceptual, creando así una nueva estructura conceptual. Luego, en la fase de generación, un intérprete de reglas recorría los nodos de la estructura conceptual y los movía a una ventana de generación, donde se verificaban relaciones de concurrencia entre las palabras, para luego aplicar unas reglas de generación que producían la traducción final.

UNL

Universal Networking Language (UNL), es el proyecto interlingua de mayor envergadura que se realiza actualmente. Inició en 1996, como un sistema de TA multilingüe que contaba con 15 idiomas. También, se proponía brindar a los computadores una herramienta para el procesamiento del conocimiento y no estaba restringido a un dominio específico. Tres componentes principales conforman el sistema: las estructuras del lenguaje, el software para procesar estas estructuras y las herramientas para mantenerlas y procesarlas. Las estructuras, se dividen en aquellas que son independientes del lenguaje (conceptos en interlingua almacenados en una base de datos de conocimientos *UNLKB*) y las que son dependientes (herramientas para el análisis y generación del *LN*). Las estructuras dependientes, al igual que el software para el procesamiento del lenguaje, se almacenan en cada uno de los servidores del lenguaje a los cuales se puede acceder a través de Internet. Por otra parte, las herramientas para producir documentos *UNL* se pueden utilizar en un computador personal y funcionan a través de Internet [17, 22]. El sistema, funciona de la siguiente manera. Un analizador de lenguaje, llamado *Enconverter*, descompone las oraciones en *LN*, de izquierda a derecha, en morfemas y aquellos que se encuentren en el diccionario se convierten en morfemas candidatos. Luego, se le aplican ciertas reglas a estos morfemas con el fin de construir el árbol sintáctico y la red semántica de la oración. Este proceso, continúa hasta obtener la red semántica expresada en formato *UNL* [22]. Si la entrada al sistema es un texto anotado o etiquetado [23] (aquellos que muestran explícitamente las relaciones semánticas entre las palabras), el sistema usa el *Universal Parser* [24] (UP), el cual genera expresiones en UNL sin necesidad de información gramática de la lengua. Posteriormente, se usa el *UNL Verifier* para verificar que la expresión *UNL* sea correcta tanto a nivel semántico como sintáctico y léxico [22].

Una vez se obtiene la expresión *UNL*, el sistema usa el *DeConverter* para generar la oración en la lengua meta. Esta herramienta, convierte la expresión *UNL* en un hipérgrafo, denominado red de nodos, donde el nodo raíz representa el predicado principal de la oración. A continuación, se aplican las reglas de generación, las cuales producen una lista de palabras en la lengua de destino y cuyo orden se determina aplicando reglas de sintaxis [22]. Sin embargo, *UNL* no es sólo un sistema de *TA* sino, además, un lenguaje artificial que permite una *RC* apropiada, por medio de un formato que se procura convertir en un estándar para el área de *PLN* [17, 25]. Este lenguaje, expresa las oraciones a través de un hipérgrafo que utiliza tres elementos principales: *universal words* (*UWs*) o palabras universales, relaciones y atributos. Las *UWs* constituyen el vocabulario del lenguaje *UNL* y se consideran los elementos léxicos básicos de *UNL*. Estos elementos, se derivaron del vocabulario del inglés, pero se modificaron a través del uso de redes semánticas para eliminar la ambigüedad de las palabras en *UNL* [17]. Estas palabras, expresan conceptos del *LN* y se definen según el tipo de relación semántica que sostenga con otros conceptos. Así, existen cuatro tipos de conceptos: los nominales, los verbales, los adjetivales y los adverbiales [26, 27]. Las relaciones, expresan el tipo de conexión semántica que sostienen los conceptos entre sí. En *UNL*, existen 46 relaciones, las cuales se seleccionan dependiendo del cumplimiento de condiciones de necesidad (debe existir todo el conocimiento para relacionar dos o más conceptos) o suficiencia (se puede comprender el rol de cada concepto con sólo referir la relación) [26]. Las relaciones, pueden ser argumentativas (agente, objeto, meta), circunstanciales (tiempo, lugar, propósito), lógicas (conjunción, disyunción), entre otras [17]. Por último, los atributos expresan toda la información semántica que resulta de la flexión morfológica, de los elementos de la frase y, en general, describen características de los conceptos. Los atributos, se dividen en ocho categorías y pueden expresar, por ejemplo, la pluralidad de un objeto (@pl.), el tiempo de un verbo (@past.), etc. [17], [26].

Procesamiento de lenguaje natural

La meta en esta área de investigación, es lograr que los computadores procesen los textos por su información semántica y no como un archivo binario [28]. Algunos proyectos que utilizan las interlinguas como apoyo para generar soluciones en esta área, se presentan a continuación.

Desambiguación de palabras

El problema de la desambiguación de palabras, consiste en elegir el significado correcto de una palabra dependiendo del contexto en el que ésta se encuentra. Las interlinguas, pueden ser herramientas útiles para la solución de este problema, ya que la RC que generan de un texto involucra análisis de tipo sintáctico, semántico y morfológico que facilita esta tarea [11, 29]. Existen diferentes formas de usar estos análisis. Algunos autores, utilizan lenguajes controlados y una serie de reglas heurísticas para reducir la ambigüedad de un texto. Además, anotan el texto utilizando etiquetas *Standard Generalized Markup Language* (SGML). Luego, generan la desambiguación trayendo conceptos semánticos desde la sintaxis usando *lexical mapping rules* y comparan la información extraída con un modelo del dominio para determinar los roles semánticos, usando reglas de interpretación semántica [30]. Sin embargo, otros autores consideran que intentar resolver el problema de la ambigüedad a través de reglas, no es suficiente para llegar a una solución real [26], e impulsan otros análisis tales como la frecuencia de las palabras en un texto, colocaciones, contextos semánticos, restricciones de selección y señales sintácticas. Así, se logran encontrar no sólo roles semánticos sino, también, si la palabra forma parte de colocaciones, de asociaciones de palabras o si es morfológicamente aceptable.

Resolución de anáforas

La anáfora es fenómeno lingüístico que se presenta cuando ciertas palabras recogen el significado de otras [13]. Según Hirst: "*anaphora, in discourse, is a device for making an abbreviated reference (containing fewer bits of disambigua-*

ting information, rather than being lexically or phonetically shorter) to some entity (or entities)" [31]. Existen diferentes tipos de anáforas según su posición en el texto, su categoría gramatical, etc. [32, 33]. La resolución de anáforas, consiste en la identificación de las palabras (generalmente pronombres, en cuyo caso se denomina *anáfora pronominal*) que hacen referencia a otras que se mencionaron previamente o se mencionarán después (catáfora) [33, 34, 35]. Para resolver este problema, los investigadores emplean diversas técnicas [36, 37]. De esas técnicas, este artículo se centra en aquellas que adoptan una aproximación interlingual para su solución. Uno de tales sistemas, es *Anaphora Generation with an Interlingua Representation* (*AGIR*) [38]. *AGIR* es un sistema de traducción automática inglés-español capaz, no sólo, de resolver la anáfora, sino también de generarla en la lengua destino. Este sistema, obtuvo una precisión del 80.4% para la traducción de la anáfora pronominal del inglés al español y del 84.8% para el caso contrario. Además, el sistema acepta cualquier tipo de texto como entrada.

El análisis comienza con la obtención de información léxica y morfológica proveniente de un anotador de partes del discurso, una división del texto en oraciones y un análisis gramatical que genera una *slot structure* (SS), la cual almacena elementos de información, como los constituyentes de la oración, marcadores del discurso e información semántica y morfológica. Luego, un módulo de desambiguación de palabras trata la SS con el fin de obtener sólo un sentido para la oración.

Una vez se obtiene el sentido, el sistema *slot unification parser for anaphora resolution* (*SUPAR*) se encarga de resolver las relaciones anafóricas, hasta conseguir una nueva estructura *SS*. El antecedente correcto en cada una de las relaciones anafóricas, se elige utilizando un *método de restricciones y preferencias* [39]. Finalmente, *AGIR* genera la representación interlingual a partir de la estructura SS y, de ahí en adelante, sigue la etapa de generación. Otro sistema de *TA* que da solución al problema de resolución de anáforas desde una orientación interlingual, es *KANTOO* [40]. Este sistema, se orienta a traducir textos técnicos (recibiendo como entrada un lenguaje controlado) y traduce anáforas del inglés al español con un 97.9% de precisión y un 94.4% para el alemán. El análisis se inicia dividiendo el texto en *tokens* (palabras, números, puntuación o una etiqueta *SGML*) a los cuales se les añade información léxica. Entonces, un analizador no determinístico genera varias estructuras sintácticas válidas y un desambiguador de palabras les elimina la ambigüedad. En ese momento, se recorre el texto buscando posibles anáforas y se ejecuta el algoritmo de resolución (basado también en el método de restricciones y preferencias) para cada anáfora encontrada. Existen siete reglas de restricción, que escogen los candidatos a relaciones anafóricas, y diez reglas de preferencias, que seleccionan el candidato correcto.

Otros usos en PLN

Los sistemas de *QuestionAnswering* (*QA*), intentan identificar la respuesta exacta a una pregunta formulada en *LN* y que se debe extraer de documentos *on-line* [41]. Las interlinguas, se convierten en una alternativa para implementar estos sistemas. *UNL*, es un ejemplo de ello [14]. Para deducir la respuesta a una pregunta directa, el primer paso consiste en transformar la pregunta a una expresión *UNL*, donde la cláusula interrogativa (¿qué?, ¿quién?, etc.) es el nodo que se necesita enlazar a una respuesta. Luego, se comienza el proceso de recorrer esta red semántica, buscando las relaciones entre los conceptos y evaluando cuál de ellas da respuesta a la pregunta solicitada.

Inteligencia artificial

La Inteligencia Artificial (*IA*), es otra disciplina que se beneficia de las características de las interlinguas, ya que la información semántica que éstas recogen permite la representación de acciones, la comunicación entre agentes inteligentes, la integración de diferentes tipos de sistemas, la interacción entre diferentes lenguajes de RC, etc.

Parameterized Action Representation (*PAR*) [42], es una interlingua que permite la representación conceptual de acciones y su objetivo es la animación de agentes humanos en realidades virtuales. Las acciones a representar, incluyen cambios de estado, de posición (cinemática) y la realización de fuerzas (dinámica). Algunas variables que permiten describir la posición, son *path, manner, duration*; otras variables como *speed* y *force* sirven para describir las propiedades dinámicas. Entre las variables de estado, se cuenta con *applicability conditions* y *preparatory actions*, las cuales se deben satisfacer antes de ejecutar una acción. *Termination conditions* y *post assertions* determinan si la acción concluye o no. *PAR*, explota la idea de que los verbos semánticamente similares se pueden asociar entre sí a un padre común que captura todas sus propiedades, creando así una jerarquía de acciones. Los nodos superiores de la jerarquía, se denominan *esquemas generalizados* y representan las estructuras argumentativas y predicativas para todo un grupo de acciones subordinadas. Por otra parte, en los nodos inferiores se encuentran los *esquemas específicos*, los cuales heredan información de los nodos superiores y se pueden ejemplificar con elementos del lenguaje natural para representar acciones específicas.

Knowledge Query and Manipulation Language (*KQML*) [43], se creó con el propósito de hacer posible la comunicación entre agentes inteligentes. Esta interlingua, proporciona el nivel de abstracción necesario para que los agentes trabajen no sólo en sistemas distribuidos sino en *IA* distribuida. Además, *KQML* se considera un formato de mensajes y un conjunto de protocolos para la administración de mensajes, capaz de compartir conocimiento en tiempo de ejecución. Para los sistemas distribuidos, *KQML* brinda la posibilidad de comunicar mensajes en protocolos tales como *TCP/IP*, http y *CORBA*. El aporte en *IA* distribuida, se establece a través del lenguaje y los protocolos que los agentes inteligentes utilizan para comunicarse entre sí. Los agentes inteligentes, se comunican a través de *KQML* usando muchas características en común que este lenguaje les ofrece tales como una semántica, una sintaxis y una ontología. La mayor fortaleza de *KQML*, es ofrecer a los agentes una pragmática común a través de la cual pueden reconocer con qué agente intercambiar información, cómo contactarlo y cómo iniciar y mantener ese intercambio. Otra de las aplicaciones de las interlinguas en la *IA*, se vincula a la representación de relaciones de tiempo y eventos como es el caso de *Versatile Event Logic* (VEL) [44]. Este lenguaje, ofrece ventajas sobre otros lenguajes formales de IA, ya que puede expresar, desde diferentes perspectivas, una representación lógica de eventos y tiempo. En general, los demás lenguajes sólo son capaces de expresar un tipo de relación con el tiempo, de los tres tipos de relaciones principales: 1) Hacer referencia al tiempo, como puntos específicos que se ordenan en una relación temporal; 2) Hacer referencia a intervalos de tiempo y a las relaciones entre ellos; y 3) Usar los tiempos proposicionales para comunicar relaciones temporales entre hechos, sin hacer referencia explícita a ninguna entidad temporal. La ventaja de *VEL*, consiste en que, a través de una sola fórmula, expresa los tres tipos de relación a la vez. En cuanto a los eventos, *VEL* también representa varios tipos de análisis: 1) Considera los eventos como una transición entre estados; 2) Considera los eventos como ocurrencias dentro de un intervalo de tiempo; 3) Comprende un análisis existencial de los eventos, en el cual se asocian los eventos con los verbos y una variable implícita sobre la existencia del evento; 4) Considera los eventos como radicales o unidades sintácticas que refieren a tipos de eventos y combinan un operador de tiempo para formar proposiciones. La estructura que le permite a *VEL* modelar todos estos tipos de relaciones entre eventos y tiempo, es su ontología, la cual posee características semánticas, a diferencia de otros lenguajes formales de IA que poseían características axiomáticas. La ontología de *VEL*, asume las historias del mundo como una estructura de árbol, a la que denomina *History Structure*. Esta estructura, consta de tuplas de la forma: $H = \{S, T, <, H\}$ donde S representa un conjunto de estados del mundo, T un conjunto de puntos en el tiempo, $<$ un orden lineal sobre T, y H un conjunto de historias que son funciones de T con contradominio en S.

Web Semántica

La Web Semántica, se está diseñando con el objetivo de que una gran cantidad de aplicaciones y servicios Web puedan usar agentes inteligentes como componentes esenciales. Sin embargo, es necesario agregar una justificación a los resultados presentados por estos nuevos servicios Web, a fin de que un cliente los pueda aceptar, manipular y confiar en ellos. La interlingua *Proof Markup Language* (PML) [45] puede resolver este problema y, además, provee una base para el razonamiento híbrido (en el cual varios agentes inteligentes producen resultados cooperando unos con otros) y para la generación de explicaciones. *PML*, se puede considerar como una ontología extendida del *Web Ontology Language* (OWL), esto es, una justificación en *PML* se expresa a través de *OWL* y, por ende, es intercambiable entre servicios y clientes de la *Web* semántica que utilicen la sintaxis propia de otros lenguajes de esta área tales como: *RDF* y *XML*. Las justificaciones de PML, son el medio a través del cual se describe toda la secuencia en que se manipula la información (inferencias, operaciones de recuperación de información, procesamiento de lenguaje natural, etc) para obtener el resultado de la búsqueda. Esta secuencia se conoce como *Proof* o prueba. PML, es un componente de la infraestructura *Inference Web* (IW) que provee las herramientas necesarias para construir, mantener, intercambiar, combinar, anotar y filtrar las pruebas. Una de las herramientas es la *IWBase*, la cual es una *hiperweb* de repositorios de metainformación que contiene principalmente los conceptos de *ProvenanceElement* e *InferenceRule*. Por otra parte, una arquitectura denominada *JTP* muestra cómo usar las pruebas *PML* para llevar a cabo el razonamiento híbrido que éstas permiten hacer.

Las principales estructuras de *PML*, son *NodeSet, InferenceStep*, *Expression*, *ProvenanceElement* e *InferenceRule*. Un *NodeSet*, es conjunto de nodos provenientes de un árbol de pruebas que tienen la misma conclusión y un mismo identificador único. Un *InferenceStep*, es una clase OWL que representa una justificación para la conclusión a la cual llegó un *NodeStep*. Una *Expresión*, es una representación en PML de expresiones lógicas escrita según un lenguaje dado. Un *ProvenanceElement*, es una superclase que revela el origen de las estructuras anteriores. Finalmente, una *InferenceRule* describe las reglas aplicadas sobre las premisas que deducen las conclusiones de un *NodeSet*.

Representación del Conocimiento

La representación del conocimiento (RC), es una función inherente a todas las interlinguas. En esta revisión de la literatura correspondiente a las interlinguas, se aprecia que todas cumplen con esta función (aunque en diferentes grados de abstracción, usando diferentes ontologías, gramáticas, sintaxis, etc.) para lograr distintas metas. Por lo tanto, la RC no se considera como una aplicación en sí misma sino como un medio a través del cual se consigue un objetivo mayor.

Todas estas diferencias entre las interlinguas, dan lugar a una nueva heterogeneidad de lenguajes de RC que conduce a la incompatibilidad entre los sistemas, a un mayor esfuerzo y costo en la implementación y mantenimiento de los mismos. Estos, son problemas típicos de la falta de estandarización que se pueden solucionar a través del desarrollo de un estándar. Para este objetivo, se desarrolló y propuso *Knowledge Interchange Format* (KIF) [46, 47]. KIF, es un lenguaje formal orientado a los computadores que permite a los sistemas intercambiar conocimiento. KIF, posee una semántica declarativa que facilita la comprensión del significado de las expresiones, sin necesidad de un intérprete que las manipule. Además, KIF incluye una lógica comprensible capaz de expresar las oraciones en cálculo de predicados, la capacidad de representar tanto el conocimiento de diferentes ontologías como reglas de razonamiento no monotónico (a través de las cuales se pueden deducir conclusiones en ausencia del conocimiento de la base de datos) y una definición precisa de objetos, funciones y relaciones. El alfabeto de *KIF*, consta de 128 caracteres del código *ASCII*. Sin embargo, una cadena de caracteres *ASCII* se considera una expresión *KIF*, si un lector del lenguaje declarativo *LISP* puede procesarla y si, además, la estructura producida por el lector es una expresión estructurada de *KIF*. Esta cercanía

a *LISP*, se debe a que *KIF* se creó utilizando ese lenguaje y, por tanto, heredó su sintaxis. La unidad básica de la sintaxis de *KIF*, es la *palabra*. A través de una palabra, es posible definir una variable, una secuencia de variables, constantes básicas (números, caracteres, cadenas de caracteres) e, incluso, una *expresión*. Las expresiones complejas, son secuencias finitas de expresiones y, entre las principales, se cuentan los *términos*, las *proposiciones*, las *reglas* y las *definiciones*. Estas expresiones, denotan objetos, expresan hechos del mundo y pasos de inferencia, definen constantes no básicas (aquellas que el usuario crea) y se crean por medio de unos *operadores*. Finalmente, se define una *forma* como una proposición, una regla o una definición y el conjunto de formas constituye *la base de conocimiento*. Así, se conforma la estructura básica de KIF.

Ingeniería de Requisitos

Las aplicaciones de las interlinguas, también se extienden a la representación de diagramas conceptuales de *UML* (*Unified Modeling Language*). Un ejemplo de ello es el *Klagenfurt Conceptual Predesign Model* (*KCPM*) [48], cuyo objetivo es mejorar el proceso de adquisición de requisitos en el desarrollo de software, a través de una participación más activa de los usuarios finales. Debido a que el *LN* es demasiado ambiguo y los modelos conceptuales son, en ocasiones, demasiado complejos para integrar activamente los aportes del usuario final, los autores argumentan sobre las ventajas del uso de una interlingua [49] para obtener automáticamente los esquemas conceptuales de *UML*. A través de *KCPM*, se pueden derivar, desde cualquier texto escrito en LN (puede utilizar un lenguaje informal u oraciones estructuradas), las características estáticas y dinámicas de los diagramas de *UML*. Por lo tanto, es posible utilizar *KCPM* para generar otro tipo de diagramas conceptuales más complejos tales como el diagrama de actividades, el diagrama de estados, el diagrama de secuencia y el de casos de uso. La estructura de *KCPM*, consta de un conjunto de nociones de modelado. Entre estas nociones están: *thing-type, connection-type, cooperation-type, operation-type, pre/post-conditions*. La noción *thing-type*, es una generalización del concepto de clase y atributo. De esta forma, es posible incluir dentro de esta categoría los conceptos de *cliente* y *nombre de cliente*. La noción *connection-type*, en cambio, representa las relaciones entre *thing-types* y, en la mayoría de los casos, corresponde a los verbos o frases preposicionales del texto. *Operation-type*, modela aquellos servicios que se invocan vía mensajes y se asemeja, así, a los conceptos de método y operación de otros modelos. Para finalizar, cuando los actores emplean algunas operaciones en ciertas circunstancias (*pre-conditions*), crean nuevas circunstancias (*post-conditions*), y este proceso se captura en una *cooperation-type*. Es decir, una cooperación es un paso elemental de algún proceso del negocio, en el cual varios actores contribuyen ejecutando una operación.

Análisis de las interlinguas

Las interlinguas, son lenguajes semánticos altamente expresivos que permiten comunicar diferentes tipos de lenguajes o sistemas entre sí. Algunas de las ventajas de utilizar estos canales son: la agilización en la comunicación, la disminución en los costos de implementación de nuevos sistemas, la estandarización y la reutilización de componentes, la facilidad para el procesamiento de textos y la extracción de información de los mismos e incluso, las inferencias que algunas interlinguas posibilitan. Todas estas ventajas, se derivan de su capacidad para representar el conocimiento. Sin embargo, la multifuncionalidad de estos lenguajes se convierte en su mayor debilidad ya que, a medida que incorporan más funciones o número de lenguajes a comunicar, mayor será el grado de abstracción que deberán tener y, por ende, se tornan más complejas las interlinguas. En el caso de *TA*, se evidencian claramente estas ventajas y debilidades, ya que los *LNs* son inexactos, ambiguos y difieren mucho entre sí. Por esto, la definición de una interlingua capaz de representar todos los *LNs* de manera que exista completa independencia de cualquier *LN* y, al mismo tiempo, tenga la expresividad suficien-

te, no es posible hasta el momento y fue la causa por la cual este método de traducción se abandonó por mucho tiempo. Esta dificultad, además, es un factor que restringe el número de lenguas a traducir porque, entre más características gramaticales y léxicas compartan las lenguas, más fácil resulta escoger un léxico o una gramática para la interlingua y, además, a menor número de ellas, las representaciones profundas son más manejables y expresivas semánticamente. Por otro lado, Moreno [9], citando a Whitelock, comenta que implementar una interlingua con una representación tan profunda, puede no ser beneficioso, ya que ésta podría no especificar valores para una determinada lengua. Sin embargo, las interlinguas ofrecen muchas otras ventajas sobre el método de transferencia, que es otra técnica de traducción. El método de transferencia, produce una representación del texto orientada hacia la lengua meta y, aunque se logran traducciones de mejor calidad, resulta ser impracticable para sistemas multilingües donde el número de lenguas (n) sea mayor que tres, porque habría que implementar n(n-1) componentes. En cambio, debido a que la representación interlingual es común a todas las lenguas a traducir, sólo necesita implementar 2n componentes, lo cual permite incluir un mayor número de lenguas. Adicionalmente, la disminución de costos en el método de interlingua es considerable, teniendo en cuenta que los mantiene dentro de una ecuación lineal, a diferencia de los costos del método de transferencia que se elevan al cuadrado. Otra de las ventajas de las interlinguas, es que agilizan el proceso de traducción y promueven la reutilización, porque para traducir sólo necesitan analizar el texto una vez y, con la misma representación, pueden generar diferentes textos metas. Por ejemplo, si se debe traducir un texto en inglés a varias lenguas, con el método de transferencia se debe analizar tantas veces como número de lenguas haya. En cambio, utilizando una interlingua sólo es necesario analizar el texto original una vez y, empleando la misma representación, se pueden generar la traducción a cualquier otra lengua. Como ventaja adicional de las interlinguas, se promueve la creación de ontologías bien estructuradas, a través de las cuales es posible separar el conocimiento de la lengua de origen y de la lengua de destino. Estas ontologías de conceptos, son herramientas importantes para la desambiguación del sentido del texto. Sin embargo, la inclusión de estas ontologías también supone problemas de diseño, implementación, manejo y mantenimiento. En el área de IA, las interlinguas sirven para comunicar diferentes tipos de sistemas y para apoyar la comunicación entre diferentes tipos de lenguajes de computador (que, a diferencia de los *LNs*, son menos ambiguos) y, por tanto, su definición es más precisa. Por esta razón, las interlinguas de *IA* son lenguajes más formales a través de los cuales es posible hacer inferencias. Algunas de estas interlinguas, intentan ser estándares en esta área, especialmente en *RC*, como es el caso de *KIF*. Sin embargo, algunos autores [50] consideran que es prematuro promover este tipo de iniciativas, ya que esta área aún pertenece al ámbito investigativo y se limitarían así los componentes a utilizar, los lenguajes y las metodologías.

Conclusiones

El desarrollo de interlinguas como lenguajes de *RC*, contribuye a la solución de diversos problemas en las áreas de *PLN* e *IA*, tales como la *TA*, la resolución de anáforas, la desambiguación de palabras, la generación de respuestas y la inferencia de información, entre otras. Estos lenguajes, impulsan y mejoran la creación de bases de conocimiento, de lexicones, de elementos de análisis sintáctico, semántico y léxico que acercan cada día vez más al hombre y la máquina, ya que permiten que ésta procese el *LN* y lo utilice para comunicarse. A pesar de la complejidad para definir este tipo de lenguajes, los beneficios que las interlinguas proporcionan demuestran el valor de las mismas. La investigación de las interlinguas puede continuar en cualquiera de las áreas que se mencionaron en este artículo y se pueden utilizar para llevar a cabo diferentes proyectos que exploten todos los recursos semánticos que estos lenguajes suministran (ontologías, lexicones, etc.). Sin embargo, aún hay muchas áreas emergentes dónde el potencial de estos lenguajes no se emplea, tales como minería de textos, generación de resúmenes y bibliotecas digitales, entre otros.

Referencias

1. J. Hutchins. *Machine Translation: General Overview*. The Oxford Handbook of Computational Linguistics. R. Mitkov (Ed.). Ed. University Press. Oxford. 2003. pp. 501-511.
2. J. Cardeñosa, A. Gelbukh, E. Tovar. "Universal Networking Language: Advances in Theory and Applications". *Research on Computer Science.* Vol. 12. 2005. pp. 1-443.
3. R. Hausser. *Foundations of Computational Linguistics: Human-computer communication in Natural Language*. Springer-Verlag. Berlín. 2001. pp. 45-49.
4. L. Zamenhof. *Fundamento de Esperanto*. Ed. Printemps. Varsovia. 1905. pp. 1-315.
5. J. Cardeñosa, E. Tovar, C. Gallardo. "El sistema UNL – Universal Networking Language". *Procesamiento del Lenguaje Natural.* Vol. 29. 2002. pp. 285-286.
6. H. Uchida. "ATLAS-II: A Machine Translation System Using Conceptual Structure as an Interlingua". *Proceedings of 2nd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*. Pittsburgh. USA. 1989. Vol. 1. pp. 150-160.
7. K. Muraki. "PIVOT: Two-phase machine translation system". *Proceeding of the School Machine Translation System Summit*. Japan. 1989. Vol. 1. pp. 113-115.
8. J. Hutchins, H. Somers. *An Introduction to Machine Translation*. Ed. Academic Press Limited. London. 1992. pp. 5-311.
9. A. Moreno. "Diseño e implementación de un lexicón computacional para lexicografía y traducción automática". *Estudios de lingüística Española*. Publicación electrónica en línea http://elies.rediris.es/elies9/. Vol. 9. 2000. Consultada el 12 de octubre de 2007.
10. E. Nyberg, T. Mitamura. "The KANT System: Fast, Accurate, High-Quality Translation in Practical Domains". *Proceedings of COLING-92: 15th International Conference on Computational Linguistics*. Nantes. France. 1992. pp. 1069-1073.
11. S. Beale, S. Nirenburg, K. Mahesh. "Semantic Analysis in the Mikrokosmos Machine Translation Project". *Proceedings of the 2nd Symposium on Natural Language Processing*. Bangkok. Thailand. 1995. pp. 297-307.
12. B. Dorr, E. Hovy, L. Levin. *Machine Translation. Interlingual Methods*. K. Brown (editor) Encyclopedia of Language and Linguistics 2ª ed. Ed. Elsevier. Oxford. 2006. pp.615-632.
13. Real Academia Española. En línea: http://www.rae.es. Consultada el 11 de octubre de 2007.
14. J. Cardeñosa, C. Gallardo, L. Iraloa. "Interlinguas: A classical Approach for the Semantic Web. A practical case". *Proceedings of the 5th Mexican International Conference on Artificial Intelligence*. Apizaco. México. 2006. Vol. 1. pp. 932-942.
15. C. Hockett. *The problem of universals in Language.* J. Greenberg (editor). Universals of Language. Ed. MIT Press. Cambridge. 2ª ed. 1966. pp. 1-28.
16. R. Langacker. *Fundamentals of Linguistic analysis*. Ed Hartcourt. Sidney. Australia. 1972. pp. 5-372.
17. J. Cardeñosa, C. Gallardo, E. Tovar. "Standardization of the generation process in a multilingual environment". *Research on Computer Science*. Vol. 12. 2005. pp. 10-24.
18. S. Nirenburg, V. Raskin, A. Tucker. "Interlingua Design for TRANSLATOR". *Proceedings of Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*. New York. USA. 1985. pp. 224-244.
19. M. Pérez. "Explotación de los corpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento". *Estudios de Lingüística Española*. Publicación Electrónica en línea: http://elies.rediris.es/elies18/. Vol.18. 2002. Consultada el 12 de octubre de 2007.
20. H. Somers. "Machine Translation: Latest Developments". R. Mitkov (Ed.). *The Oxford Handbook of Computational Linguistics*. Ed. University Press. Oxford. 2003. pp. 512-528.
21. C. Boitet. "A Rationale for Using UNL as an Interlingua and More in Various Domain". *Research on Computer Science*. Vol. 12. 2005. pp. 3-9.
22. Universal Networking Digital Language Foundation. http://www.undl.org/index.php?option=com_content&task=view&id=26&Itemid=57 Consultada el 11 de octubre de 2007.
23. M. Zhu, H. Uchida. UNL Annotation. 2003. En: http://www.undl.org/unlsys/uparser/UNLA.pdf. Consultada el 25 de noviembre de 2007.
24. M. Zhu, H. Uchida. Universal Parser. 2003. En: http://www.undl.org/unlsys/uparser/UP.htm. Consultada el 25 de noviembre de 2007.
25. C. Boitet, G. Sérasset. "On UNL as the future "html of the linguistic content" & the reuse of existing NLP components in UNL-related applications with the example of a UNL-French deconverter". *Proceedings of COLING 2000: 18th International Conference on*

Computational Linguistics. Saarbrucken. Germany. 2000. Vol. 1. pp.768-774.

26. Universal Networking Language Specifications Version 2005. http://www.undl.org/unlsys/unl/unl2005-e2006/. Consultada el 11 de octubre de 2007.
27. I. Boguslavsky. "Some Controversial Issues of UNL: Linguistic Aspects". *Research on Computer Science*. Vol. 12. 2005. pp. 77-100.
28. A. Gelbukh, G. Sidorov. *Procesamiento automático del español con enfoque en recursos léxicos grandes*. Instituto Politécnico Nacional. México. D.F. 2006. pp. 13-58.
29. S. McRoy. "Using Multiple Knowledge Sources for Word Sense Discrimination". *Computational Linguistics*. Vol. 18. 1992. pp. 1-30.
30. K. Baker, A. Franz, P. Jordan, T. Mitamura, E. Nyberg. "Copying with ambiguity in a Large-Scale Machine Translation System". *Proceedings of COLING*. Japan. 1994. pp. 90-94.
31. G. Hirst. "Anaphora in Natural Language Understanding: A survey". *Lecture Notes in Computer Science*. 1981. pp. 1-128.
32. R. Mitkov. *Anaphora Resolution*. London. Longman. 2002. Capítulo 1.
33. I. Aduriz, K. Ceberio, A. Díaz. "Pronominal anaphora in Basque: annotation of a real corpus". *XXII Congreso de la SEPLN*. Zaragoza. España. 2006. pp. 99-104.
34. R. Mitkov, S. Choi, R. Sharp. "Anaphora Resolution in Machine Translation". *Proceedings of The Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*. Leuven. Bélgica. 1995. Vol . 1. pp. 87-94.
35. R. Mitkov. *Anaphora Resolution: The state of the art*. Technical Report. University of Wolverhampton. 1999. pp. 1-34. En: http://citeseer.ist.psu.edu/352217.html. Consultada el 25 de noviembre de 2007.
36. T. Deoskar. "Techniques for Anaphora Resolution: A survey". Computer Science. Cornell University. 2004. En: http://www.cs.cornell.edu/courses/cs674/2005sp/projects/tejaswini-deoskar.doc. Consultada el 25 de noviembre de 2007.
37. N. Ge, J. Hale, E. Charniak. "A statistical Approach to Anaphora Resolution". *Proceedings of the Sixth Workshop on Very Large Corpora*. Montreal. Canadá. 1998. pp. 161-170.
38. J. Peral, A. Ferrández. "Translation of Pronominal Anaphora between English and Spanish: Discrepancies and Evaluation". *Journal of Artificial Intelligence Research*. Vol. 18. 2003. pp. 117-147.
39. A. Férrandez, M. Palomar, L. Moreno. "An empirical Approach to Spanish Anaphora Resolution". *Machine Translation*. Vol. 14. 1999. pp. 191-216.
40. T. Mitamura, E. Nyberg, E. Torrejon, D. Svoboda, A. Brunner, K. Baker. "Pronominal Anaphora Resolution in the KANTOO Multilingual Machine Translation System". *Proceedings of 9th International Conference on Theoretical and Methodological Issues in Machine Translation*. Japan. 2002. Vol. 1. pp. 115-124.
41. S. Harabagiu, D. Moldovan. "Question Answering". *The Oxford Handbook of Computational Linguistics*. R. Mitkov (editor). Ed. University Press. Oxford. 2003. pp.560-582.
42. K. Kipper, M. Palmer. "Representation of Actions as an Interlingua". *Proceedings of the Third Workshop on Applied Interlinguas, held in conjunction with ANLP-NAACL*. Seattle. USA. 2000. Vol. 1. pp. 12-17.
43. T. Finin, R. Fritzson, D. McKay, R. McEntire. "KQML as an Agent Communication Language". *Proceedings of the Third International Conference on Information and Knowledge Management*. Gaithersburg. Maryland. USA. 1994. Vol. 1. pp. 456-463.
44. B. Bennett, A. Galton. "A unifying semantics for time and events". *Artificial Intelligence*. Vol. 153. 2004. pp. 13-48.
45. P. Pinheiro, D. McGuinness, R. Fikes. "A proof markup language for Semantic Web services". *Information Systems*. Vol. 31. 2006. pp. 381-395.
46. M. Genesereth, R. Fikes. *Knowledge Interchange Format, version 3.0 Reference Manual*. Technical Report Logic-92-1. Computer Science Department, Stanford University. Palo Alto. CA. USA. 1992. pp. 1-68.
47. R. Patil, R. Fikes, P. Patel Schneider, D. McKay, T. Finin, T. Gruber, R. Neches. "The DARPA Knowledge Sharing Effort: Progress Report". *Proceedings of the Third International Conference on Principles of Knowledge Representation and reasoning*. Cambridge, MA, USA. 1992. Vol. 1. pp. 777-788.
48. M. Ginsberg. "Knowledge Interchange Format: The KIF of Death". *AI Magazine*. Vol. 12. 1991. pp. 57-63.
49. G. Fliedl, C. Kop, H. Mayr, A. Salbrechter, J. Vöhringer, G. Weber, C. Winkler. "Deriving static and dinamyc concepts from software requirements using sophisticated tagging". *Data & Knowledge Engineering*. Vol. 61. 2007. pp. 433-448.
50. G. Fliedl, C. Kop, H. Mayr. "From textual scenarios to a conceptual schema". *Data & Knowledge Engineering*. Vol. 55. 2005. pp. 20-37.