NOTAS

LINGÜÍSTICA COMPUTACIONAL: UN ESBOZO

Andrés Domínguez Burgos
computerlinguist@gmx.de

El siguiente artículo tiene como objetivo presentar a los hispanoparlantes interesados en la industria de las tecnologías del lenguaje una breve guía de lo que es la lingüística computacional.

La lingüística computacional (LC) es una ciencia interdisciplinaria que se ubica entre la lingüística y la informática, con énfasis en la lingüística. Su fin es la elaboración de modelos computacionales que reproduzcan uno o más aspectos del lenguaje humano. Dos áreas aledañas a la LC son el procesamiento del habla realizado por parte de la informática y el reconocimiento de voz desarrollado por la ingeniería eléctrica. Las fronteras entre estas áreas del conocimiento son bastante difusas y pueden ser consideradas como meras convenciones.

La LC no es ámbito exclusivo de centros académicos. A decir verdad, hoy en día la LC está siendo impulsada en gran medida por la industria privada. Las empresas dedicadas a la informática han reconocido desde hace tiempo que el procesamiento automático del lenguaje humano constituirá uno de los principales campos de desarrollo en las próximas décadas.

Algunos de los programas en los que trabajan los lingüistas computacionales se usan para:

- elaborar modelos de teorías lingüísticas;
- enseñar idiomas extranjeros;
- corregir la ortografía y la sintaxis de textos en un idioma dado;
- reconocer la voz humana y procesar la información contenida en frases pronunciadas naturalmente por cualquier persona;
- crear sistemas expertos que respalden la labor de especialistas en un área dada;
- desarrollar juegos digitales que usen de una u otra forma el lenguaje humano;
- producir traducciones automáticas de textos o ayudar a traductores humanos en su trabajo;
- generar voz artificial con alto grado de naturalidad para la transmisión de información por teléfono, etc.

Un mismo programa puede tratar varios de estos puntos. Por ejemplo, actualmente la Unión Europea está financiando un proyecto (Euromobil) para producir un intérprete virtual, es decir, un sistema que oye un mensaje en un idioma A y que produce un mensaje, también oral, con un contenido comunicacional equivalente en un idioma B. Aquí se conjugan tecnologías de reconocimiento y producción automáticos de voz con traducción automática.

1. Breve historia de la lingüística computacional

El término "lingüística computacional" comenzó a usarse en los años sesenta, pero ya a finales de la Segunda Guerra Mundial se había estado trabajando en este campo. De hecho, uno de los primeros usos que se les dio a las computadoras fue en el área del procesamiento del lenguaje humano. La primera demostración de un sistema de traducción automática tuvo lugar en 1954. En pocos años diversos equipos en Estados Unidos y la Unión Soviética trabajaban en diversos proyectos para la creación de programas de traducción entre el inglés y el ruso. Los militares y los servicios de inteligencia de ambos países tenían un interés particular en dichos proyectos y se convirtieron en los principales inversionistas en este campo durante largo tiempo.

En los años 40 y 50 se produjeron grandes avances en dos áreas que resultarían claves para la evolución de las tecnologías de procesamiento del lenguaje humano: la teoría de los autómatas y los modelos probabilísticos o de teoría de la información. La teoría de los autómatas proviene de los trabajos del matemático británico Alan Turing, uno de los padres de la computadora. El estadounidense Claude Shannon (1948) aplicó la teoría de la probabilidad de procesos de Markov para desarrollar autómatas que procesaran el lenguaje humano. Asimismo, contribuyó a la evolución de la LC con sus ideas sobre el canal de ruido y decodificación del lenguaje.

Noam Chomsky, basándose en parte en los trabajos de Shannon, consideró la utilización de máquinas de estados finitos para describir posibles gramáticas. Definió un lenguaje de estados finitos como un lenguaje generado por una gramática limitada que llamó "de estados finitos". Sus comentarios críticos sobre las limitaciones de los modelos de Markov para describir gramáticas humanas contribuyeron a que cesase gran parte de la investigación en dicho campo. Aun así, sus trabajos en el terreno de la teoría de lenguajes formales propulsaron en medida significativa el desarrollo de una teoría de los lenguajes de programación.

A finales de los años 50, las investigaciones fueron concentrándose en dos áreas principales: un campo simbólico y uno estocástico. El enfoque

simbólico produjo a su vez dos corrientes importantes: una liderizada por Chomsky y científicos de la computación y de la lingüística formal interesados en el análisis sintáctico, y otra interesada en la inteligencia artificial. En esta última área se destacaron, entre otros, Marvin Minsky y, nuevamente, Claude Shannon.

El campo estocástico estuvo representado principalmente por los ingenieros eléctricos, que trabajaban con estadísticas y probabilidades. De sus investigaciones salió el método de Bayes para reconocimiento óptico de caracteres y para probar la autoría de textos (aunque la matemática de Bayes se remonta al siglo XVIII). Ya en los años 40 los ingenieros habían desarrollado el espectrógrafo de sonido, que permite el análisis de las ondas sonoras, y, en los años 50, obtuvieron los primeros reconocedores artificiales de voz.

Los años de euforia prematura en la industria de la traducción automática tuvieron un fin abrupto con la aparición en 1965 del informe ALPAC de la National Academy of Science, donde se describían los limitados resultados obtenidos hasta ese momento. ALPAC resultó un duro golpe para los que trabajaban en el campo de la traducción automática: el financiamiento fue reducido de manera drástica y durante mucho tiempo la investigación se limitó a unos cuantos proyectos, principalmente en Europa occidental y Asia. La Association for Machine Translation and Computational Linguistics incluso decidió llamarse sencillamente "Association for Computational Linguistics". Otras áreas del procesamiento automático del lenguaje pasaron a ocupar el interés principal.

El corpus Brown de inglés americano fue el primer corpus de gran envergadura y el que motivó en los años 60 diversas investigaciones en la lingüística de corpus.

ELIZA fue uno de los primeros programas con la capacidad de llevar a cabo conversaciones limitadas con los usuarios. El sistema era relativamente sencillo y se basaba en el reconocimiento de expresiones regulares, pero llegaba a crear la impresión en muchas personas de que poseía cierto grado de inteligencia (aunque realmente lo que hacía era ante todo modificar los enunciados del usuario y presentarlos como lo haría un psicólogo). M. Ross Quillian comenzó a trabajar en redes semánticas poco después y su trabajo fue continuado por otros científicos como Roger Shank y Yorick Wilks. La gramática de casos del lingüista Charles Fillmore fue usada para complementar estas redes.

En los años 70 los trabajos estocásticos condujeron al desarrollo de algoritmos de reconocimiento de voz cada vez más avanzados en centros de investigación como los de IBM y los Laboratorios de AT&T Bell.

El interés por las posibilidades de la lógica condujo al desarrollo de PROLOG, un lenguaje de programación predicativo muy usado en diversas áreas de la Inteligencia Artificial y de la LC. Algunas corrientes lingüísticas como la GRAMÁTICA FUNCIONAl y la GRAMÁTICA LÉXICA FUNCIONAL contribuyeron a que se utilizaran cada vez más mecanismos de unificación. En esa época, Terry Winograd presentó el sistema SHRDLU, un programa de reconocimiento del lenguaje natural que simulaba un robot capaz de manipular bloques imaginarios.

A finales de los 80 volvió la corriente empírica, que había sido desacreditada por tanto tiempo, en gran parte por las críticas de Chomsky y otros científicos simbolistas. Los modelos probabilísticos dejaron de ser dominio primordial de los ingenieros en el área de reconocimiento de voz y comenzaron a ser utilizados con una creciente frecuencia para el análisis morfológico y sintáctico, para la traducción automática y para muchas otras áreas.

Los años 90 han visto la revolución de Internet y una consecuente necesidad de perfeccionar las tecnologías de procesamiento automático del lenguaje. Actualmente en todo el mundo industrializado numerosas empresas y centros académicos trabajan en el área. La LC aún está en su infancia, pero su desarrollo es cada vez más acelerado.

2. ÁREAS DE LA LC.

Todo aspecto del lenguaje humano puede ser de interés para la LC. Se trabaja en desarrollar aplicaciones para el análisis automático de la fonética, la fonología, la morfología, la sintaxis, la semántica y la pragmática. La generación del lenguaje puede implicar desde métodos para transformar conceptos complejos en representaciones semánticas fácilmente procesables por máquinas, hasta la transformación de un texto en un lenguaje concreto y con convenciones muy particulares en una voz de apariencia humana. Más allá, los científicos procuran desarrollar sistemas que posibiliten el diálogo entre humanos que usan idiomas diferentes (traducción automática) o entre humanos y máquinas (sistemas de diálogo, sistemas expertos). Algunas de las áreas de trabajo para la creación de programas que procesan el lenguaje humano son descritas brevemente a continuación:

2.1. Etiquetamiento morfológico o Tagging:

Se entiende por *etiquetamiento morfológico* o *tagging* en la LC el análisis morfológico automático de las palabras que componen una frase

dada. Sin tener en cuenta el resto de la oración, la palabra *como*, por ejemplo, puede ser un verbo, una conjunción o un adverbio. Datos de carácter sintáctico y/o semántico son con frecuencia necesarios si se quiere obtener una sola interpretación posible de la morfología de una palabra dada en una oración. A veces, pese a todos los análisis posibles, quedan ambigüedades que deben resolverse mediante criterios estadísticos o por procesamientos de razonamiento que tienen en cuenta un conocimiento enciclopédico almacenado en algún banco de datos. Si las ambigüedades no son resueltas en el *tagging* mismo, pueden ser analizadas en una posterior etapa de análisis sintáctico-semántico.

2.2. Análisis Sintáctico o Parsing:

Parsing en la LC es el análisis automático de una oración dada. Tradicionalmente se consideraba en la LC el proceso de *parsing* como limitado al análisis de la sintaxis de una oración. En la Inteligencia Artificial (AI en inglés) se incluyen procesos de interpretación semántica dentro del concepto de *parsing*.

Los tipos de algoritmos para análisis sintáctico pueden ser clasificados *grosso modo* según sigan un enfoque *bottom-up* o uno *top-down* o según sean direccionales o no. En un enfoque *bottom-up* se procede a utilizar los símbolos terminales para ir reconociendo estructuras cada vez más complejas, hasta llegar a frases, y de allí a construir una oración u oraciones. El enfoque *top-down* parte de que la información que se obtiene es una oración dada, y realiza hipótesis sobre qué frases pueden constituir la oración dada y cómo están organizadas estas frases, para así hasta llegar a los elementos o símbolos terminales. En ciertas propuestas de análisis sintáctico se combinan ambos enfoques.

El análisis superficial o *shallow parsing* es un procesamiento parcial cuya única función es identificar ciertos componentes de la oración sin llegar a un análisis exhaustivo. Este tipo de análisis es suficiente para diversas aplicaciones, por ejemplo, para ayudar a un análisis morfológico o para fines de recuperación inteligente de información (*information retrieval*).

Los problemas de ambigüedad son relativamente sencillos a nivel morfológico en comparación con los que se producen a nivel sintáctico. Se usa cada vez más la estadística para resolver ambigüedades sintácticas y semánticas.

2.3. Técnicas de reconocimiento de voz y conversión de texto a voz

Un sistema de reconocimiento automático de voz (*automatic speech*

recognition o ASR) es un artefacto que transcribe de manera automática la voz humana en datos que puedan ser procesados por la computadora. El reto básico consiste en traducir una señal acústica continua en una serie de símbolos discretos, un texto que equivalga a la fiel representación de la señal en cuestión. La comprensión automática del habla (*automatic speech understanding* o ASU) intenta ir más allá: producir algún tipo de procesamiento semántico de la oración.

Los reconocedores de voz se están utilizando cada vez más en los sistemas de información de estaciones de trenes, en centrales telefónicas, en los sistemas operativos de las computadoras personales, en los teléfonos portátiles y en muchos otros casos donde se quiere automatizar la comunicación. En el reconocimiento automático de voz, la señal acústica se considera como un flujo de datos que pasa por un canal de ruido: hay que decodificar la información que está mezclada con el ruido del ambiente. Para ello, los reconocedores intentan identificar, entre un conjunto enorme de oraciones potenciales, aquella con la mayor probabilidad de ser la fuente de la señal acústica. Los reconocedores de voz se basan en modelos que representan la probabilidad de las oraciones (entendida una oración como una serie n de palabras o n-gramas dados), en modelos que representan la probabilidad de que una palabra determinada sea realizada como un grupo de fonemas dados (usualmente por medio de HMMs o *Hidden Markov Models*) y en modelos que expresan la relación probabilística entre fonemas y características acústicas o espectrales dadas (modelos de Gauss o MLPs). Entre los sistemas más avanzados de reconocimiento de voz se encuentran los de IBM, L&H, AT&T y DRAGON.

La conversión de texto a voz tiene como objetivo generar de manera automática los sonidos que produciría una persona al leer en voz alta cualquier texto. El sistema que realice esta conversión deberá estar en capacidad de identificar todos aquellos elementos que no correspondan a la codificación gráfica de la mayoría de las palabras, sino que sigan alguna convención particular, como la lectura de abreviaturas, palabras extranjeras, fórmulas matemáticas, etc. Un sistema de conversión de texto a voz deberá asimismo producir los sonidos no sólo de una manera inteligible, sino también natural. Básicamente el sistema transforma un texto recibido en una transcripción fonética del mismo que debe incluir información sobre estructura sintáctica, pausas y entonación. Los sintetizadores pueden generar los sonidos con un modelo de producción de voz, ya sea en base a un enfoque articulatorio o uno de formantes o con un modelo que imita la producción de la señal de voz, como lo hacen los sintetizadores por concatenación. Estos últimos utilizan grabaciones de un hablante real y las manejan en base a cada alófono, a difonos o trifonos. Los

sistemas de conversión de texto a voz tienen cada día más uso en la lectura de textos predefinidos o determinados por un sistema de diálogo con el usuario. Entre los sistemas de conversión de texto a voz se encuentran los comerciales de AT&T, Nuance, Babel y L&H, así como sistemas no comerciales como los desarrollados por el Instituto Politécnico de Mons (MBROLA).

2.4 Recuperación inteligente de información o Information retrieval

La recuperación digital de información es un campo muy amplio que incluye todas las formas de almacenamiento y envío digital de datos de cualquier índole. En el caso de la LC, se trata principalmente de técnicas para la extracción de datos contenidos en textos y su transmisión a los usuarios. Para ello se usan actualmente métodos de procesamiento estadísticos y simbólicos diversos. Los buscadores de Internet se basan en uno o más de estos métodos de recuperación de información.

Los métodos más primitivos consisten en cálculos de diversas frecuencias. El modelo de espacio de vectores, uno de los procedimientos estadísticos más usados en este campo, es una forma de ver los textos y las peticiones de textos determinados como vectores multidimensionales gigantescos. Los textos cuyos vectores sean más parecidos a los vectores que representan las preguntas del usuario tienen una mayor probabilidad de corresponder a la información deseada. Para búsquedas más sofisticadas se trata de desambiguar el sentido de las palabras mediante el uso combinado de estadísticas, redes semánticas, sistemas ontológicos, etc. Actualmente se trabaja en gran medida en el desarrollo de algoritmos que posibiliten el aprendizaje automático de los datos necesarios para procesar otros datos y en un empleo más sofisticado de las redes semánticas.

Los sistemas de diálogo inteligentes pueden optimizar el envío de los datos preseleccionados.

2.5 Sistemas de diálogo y sistemas expertos

Los sistemas de diálogo son básicamente sistemas que permiten la comunicación entre uno o más usuarios y la computadora. Los sistemas de diálogo pueden ser tan primitivos como las rutinas de preguntas y respuestas utilizadas para instalar ciertos programas o bien tan avanzados que pueden simular en gran medida un interlocutor humano. Una gran cantidad de los sistemas de diálogo que se están desarrollando actualmente tienen que ver con el área de planificación para compras (principalmente de boletos de viajes,

pero también para muchos otros fines). Los sistemas de diálogo pueden basarse en algoritmos complejos que funcionan sobre un plan de inferencias y otros algoritmos menos complejos basados en palabras clave.

Los sistemas expertos, un campo muy estudiado dentro de la Inteligencia Artificial, son representaciones del conocimiento de expertos en un campo dado que han sido almacenadas de forma digital. Un sistema experto está compuesto por un *software* de manipulación, una banco de datos que contiene los hechos y reglas válidas para el área de conocimiento que se representa, un componente de generación de inferencias a partir del *software* y del banco de datos, así como una interfaz con el usuario. Esta interfaz puede ser realizada en forma de un sistema de diálogo. Entre los usos más corrientes de sistemas expertos está el diagnóstico automatizado de enfermedades (generalmente como respaldo a la labor de los médicos) y la inspección mecanizada de equipos altamente sofisticados.

Un campo relacionado con estos dos es el de generación automática de textos, un proceso en el que una serie de datos no lingüísticos es transformada en lenguaje natural. Para que se realice una generación automática de textos tiene que haber una selección de contenido, del léxico adecuado para expresar dicho contenido, de una estructuración de dicho léxico en oraciones, y de oraciones en un discurso cohesivo y coherente. Entre los enfoques básicos de elaboración del discurso se encuentran el de esquemas de textos y el de planificación de relaciones retóricas. Entre las gramáticas más usadas para *software* de generación de texto se hallan la Gramática Sistémica y la Gramática de Unificación Funcional. La Teoría de Sentido-Texto de Mel'cuk también ha contribuido a la elaboración de este tipo de *software*. Entre los *softwares* de generación automática de texto se encuentran KAMP, PENMAN, BABEL, COMET, REALPRO y ERLI. La generación automática de textos variados y sin hacer uso de simples "plantillas" es uno de los campos de investigación más recientes en la lingüística computacional.

2.6. Traducción automática

Los sistemas de traducción han sido criticados por sus grandes limitaciones frente a un traductor humano. Muchas personas desconocen que la traducción es a menudo una labor ardua e interesante que requiere no sólo de una comprensión muy profunda de dos sistemas lingüísticos dados, sino también de dos culturas y técnicas de comunicación. Si aún no se puede esperar que un sistema digital reemplace a un médico o a un conductor de ambulancias, tampoco se puede esperar que un *software* realice una labor tan

efectiva como un ser humano a la hora de comprender un mensaje en un idioma dado y de trasmitirlo a uno diferente teniendo en consideración todos los elementos pragmáticos necesarios. 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.

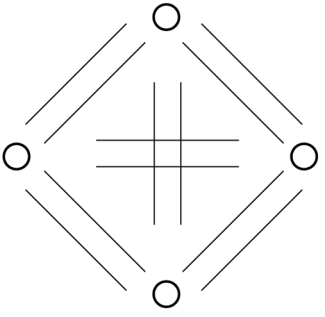
Los enfoques principales para programas de traducción automática son tres: el enfoque de traducción palabra por palabra, la traducción por transferencia y la traducción por medio de una llamada "interlingua". El enfoque de traducción palabra por palabra es el enfoque trivial, y apenas es utilizado hoy en día. Como el contenido semántico y pragmático de una palabra no se puede determinar generalmente de manera aislada, es difícil conseguir una traducción adecuada de alguna frase con el reemplazo de una palabra por otra. La descripción del mundo por parte de una comunidad lingüística siempre es diferente a la que hacen otras comunidades, por más cercanas que estén dichos grupos sociales. Estas diferencias, sumadas a las innumerables divergencias en el uso de recursos morfosintácticos en uno u otro idioma hacen que, en la gran mayoría de los casos, un simple reemplazo de una palabra del original por otra en el idioma de destino se convierta en una empresa fútil. En la mayoría de los casos es una casualidad que se pueda hablar de una traducción. Aun así, el enfoque de traducción directa puede tener ciertos usos cuando se trata de traducir lenguajes muy emparentados, como es el caso del español y del portugués, del neerlandés y del alemán o del ruso y del ucraniano.

El enfoque de transferencia es el más usado hoy día. El análisis sintáctico y semántico de las oraciones del texto origen produce árboles de estructuras que son transformados mediante una serie de reglas en árboles de estructuras que, tras diversas transformaciones a nivel sintáctico y morfológico, se convertirán en oraciones en el idioma de destino. Un ejemplo de este enfoque lo constituye SYSTRAN, disponible gratis en Internet para la traducción de textos limitados.

En el enfoque de traducción por medio de una interlingua, el programa de traducción automática analiza morfosintáctica, semántica y pragmáticamente el texto original y con esta información produce una representación intermedia en un lenguaje humano natural o artificial (como el inglés o el esperanto) o en una representación totalmente abstracta del significado. A partir de este lenguaje o de esta representación abstracta, el módulo de síntesis procede a elaborar un texto en el lenguaje de destino. Un ejemplo de este enfoque es el sistema Rosetta, de Phillips. En este sistema, se usa el esperanto como *interlingua*.

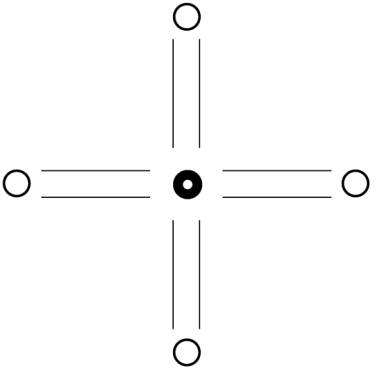
Una de las ventajas principales de usar una representación intermedia puramente abstracta es que la abstracción en forma de *interlingua* puede ser reutilizada para la traducción automática con otros idiomas. Para traducir del idioma A al idioma B mediante el enfoque de transferencia, se requiere de un módulo de análisis del idioma A al idioma B, y uno de generación del texto en B. Si se desea añadir un nuevo idioma C en el sistema de traducción, se tiene que agregar un nuevo módulo de análisis y uno de generación para cada dirección de la traducción (entre C y A y entre C y B). Si el número de idiomas de un sistema de traducción por transferencia es n, el número total de módulos es igual a n(n-1) (Ver Fig. 1). En el caso de un sistema de *interlingua*, sólo se necesita, por cada idioma natural que se agregue un solo módulo de análisis y uno de generación. Con la *interlingua*, el número total de módulos es igual a 2n (Ver Fig. 2). Una de las principales dificultades que caracteriza el enfoque de la *interlingua* es que debe ser lo suficientemente específico para representar los conceptos y relaciones que pueden expresarse con cualquier idioma natural que vaya a ser procesado por el *software* en cuestión, un objetivo al que todavía se está lejos de llegar.

Figura 1



4 idiomas
4(4-1)=12 módulos

Figura 2



4 idiomas, 1 interlingua
4∗2= 8 módulos

(Ya sea una representación semántica abstracta o algún idioma como el esperanto)

: idioma natural

: interlingua

: modulo de conversión

Actualmente se intenta combinar la traducción por transferencia o *interlingua* con otros enfoques basados en la utilización de estadísticas y de corpora de traducciones. Además, se buscan cada vez más los mecanismos que faciliten la labor del traductor humano, las llamadas Memorias de Traducción. Asimismo, los sistemas de traducción usan cada vez más un componente semántico-pragmático, y en algunos casos están comenzando a resolver anáforas y otras referencias no sólo a nivel de la oración, sino a nivel del texto.

3. El Estudio de la LC

La LC se puede estudiar en algunos centros como una carrera completa, mientras que en otros se cursa en combinación con otras asignaturas (usualmente en el área de la lingüística, la matemática o la informática) o como un curso más o menos general en el contexto de otra carrera. Las instituciones que proporcionan estudios en el área de LC aumentan de día en día. En la bibliografía se ofrece información sobre algunas de esas instituciones.

Para aquellos interesados en estudiar por su cuenta la LC, pueden usar la siguiente sección como guía inicial bastante rudimentaria.

La LC abarca *grosso modo* tres áreas de estudio: la lingüística propiamente dicha, la informática y la lingüística algorítmica. Un estudiante de LC debe tener ciertos conocimientos sobre las teorías estructuralistas, sobre las teorías generativistas clásicas, sobre la Gramática de Casos de Fillmore, sobre la *Head-Driven Phrase Grammar*, la *Functional Grammar*, etc. Aquí no hablaremos más de ese campo.

La división que se ha producido a partir de la Edad Media entre las humanidades y las ciencias naturales o exactas ha tenido en parte consecuencias fatales. Muchas personas han llegado a creer que el mundo de las palabras y el de los números son irreconciliables o al menos conflictivos. Esto ha llevado a que muchos especialistas de las llamadas ciencias naturales o exactas sientan cierto desprecio o temor por los estudios humanísticos y, a su vez, que muchos especialistas del área de humanidades sientan lo mismo por los estudios en las ciencias naturales o exactas. Una persona interesada en la LC, ya sea que venga del área de la informática o de la lingüística, debe saber apreciar tanto las palabras como los números y la lógica.

Un lingüista computacional, evidentemente, debe saber programar y tener buenos conocimientos de los algoritmos fundamentales usados en la informática. Los lenguajes de programación evolucionan cada día y una recomendación que se haga hoy sobre lenguajes de programación particular-

mente útiles o frecuentemente usados entre los lingüistas computacionales puede ser obsoleta en unos años. Hasta ahora C y C++ se han mantenido como herramientas básicas. Para programación en WINDOWS también es usado con bastante frecuencia VISUAL BASIC, una extensión del viejo BASIC o VISUAL C++, una extensión de C++. Para programas que corran en Internet, JAVA es una buena solución. Es relativamente parecido a C++, pero un programa en Java, a diferencia de uno compilado en C o C++, puede correr sin cambios en casi la totalidad de las computadoras actualmente ofrecidas en el mercado (siempre y cuando la computadora tenga instalado el intérprete Java). Tanto C++ como Java son usados primordialmente para aplicaciones, para productos finales. Para fines de investigación y para procesamiento interno de datos, se usa actualmente con mucha frecuencia lenguajes *scripts* como PERL. Para crear y hacer funcionar programas en Perl y Java, se pueden conseguir intérpretes gratis en Internet (http://www.perl.com para el intérprete de Perl y www.sun.com para el JAVA TOOLKIT y en http://www.borland.com para la versión personal del JBUILDER de Borland). Para crear programas en C++ se puede conseguir un compilador *freeware* en Internet (usualmente bajo la consigna GNU). Para la creación de *software* de carácter comercial se emplean usualmente intérpretes y ambientes de desarrollo más sofisticados (como Visual C++ para aplicaciones que funcionan en ambiente Windows).

Es primordial hacer un curso en estructuras de datos. Quienes quieran tener una impresión de los contenidos de un curso semejante puede referirse a los textos de Alfred Aho o Robert Sedgewick. En un curso de estructuras de datos, los estudiantes aprenden a transformar una solución a un problema expresada en palabras paso a paso en un algoritmo escrito en un lenguaje de programación como C. Luego aprenden a organizar datos básicos (como caracteres o números) en estructuras más complejas (como información de un estudiante o datos de un cliente o información morfosintáctica de un verbo) y a desarrollar algoritmos efectivos para clasificar y buscar dichas estructuras de datos, así como para realizar todo tipo de cálculo con las mismas.

Actualmente también es necesario manejar bancos de datos, saber qué son claves primarias, claves secundarias, cómo se evita la redundancia de la información, etc. El lenguaje estándar actual para bancos de datos es SQL. Esta es una herramienta imprescindible para todo aquel que quiera dedicarse de una manera seria al desarrollo de programas.

En el área de la matemática también son absolutamente necesarios los conocimientos de lógica (principalmente lógica declarativa y predicativa, pero también lógica modal y lógica difusa). Asimismo, es conveniente estar familiarizado con la teoría de conjuntos, funciones, autómatas, y la teoría de

lenguajes formales en general. Es igualmente útil el manejo básico de conceptos relacionados con la teoría de grafos y el álgebra lineal, particularmente para quienes quieran participar en la elaboración de programas de cierta complejidad. Finalmente, la estadística y la teoría de probabilidades son áreas de la matemática que también pueden ayudar en gran medida al científico en este campo.

Un estudio de lingüística computacional se cierra usualmente con una tesis que involucra el desarrollo de un *software* para una labor de procesamiento del lenguaje. Como actualmente casi ningún programa complejo es realizado por un solo individuo y como es imprescindible conocer cómo se organiza el trabajo entre varios programadores, es recomendable que se realice un proyecto en un equipo.

4. Referencias

La principal asociación es la Association for Computational Linguistics (ACL). Hay una rama europea, la *European Association for Computational Linguistics* y una americana, la *American Association for Computational Linguistics*. Cada dos años se celebra la *International Conference on Computational Linguistics* (COLING). Los documentos producidos en COLING y en otros eventos relacionados con las organizaciones antes mencionadas son de interés para todos aquellos que quieren seguir de cerca el desarrollo de la LC.

Otros congresos de interés para los lingüistas computacionales son la bienal de *Applied Natural Language Processing* (ANLP), el congreso de *Empirical Methods in Natural Language Processing* (EMNLAP) y la *International Conference on Spoken Language Processing* (ICSLP).

Actualmente ningún científico puede decir que posee un conocimiento profundo de todas las áreas de la lingüística computacional. La especialización ha progresado lo suficiente como para que se tengan que formar grupos de discusión particulares para las diversas facetas del procesamiento mecanizado del lenguaje. Desde el sitio de la ACL en Internet se puede acceder a diversos de estos grupos.

Daniel Jurafsky & James Martin (2000) ofrecen una buena introducción general a todas las áreas de la LC aquí mencionadas, así como una extensa bibliografía. Para conocer el uso de estadísticas y probabilidades en el procesamiento del lenguaje, se puede consultar a Christopher Manning & Hinrich Schütze (1999). Una excelente introducción a la teoría de *parsing* la constituye el manual del Prof. Peter Hellwig, así como el libro de Dick Grune & Ceriel

Jacobs (1990). El análisis sintáctico asistido por la estadística es tratado en Manning & Schütze. Para el campo del reconocimiento de voz, la introducción de Daniel Jurafsky & James Martin es una buena guía. Para un tratado más matemático, el lector puede referirse a Frederik Jenilek (1999). John Hutchins (1986 y 1992) presenta amplia información interesante sobre la traducción automática. Partee *et al* (1990) ofrece una excelente iniciación en varios aspectos de la lingüística matemática. En la bibliografía están mencionados otros libros que también pueden resultar motivadores para el lector de este artículo.

Internet, naturalmente, es una de las mejores fuentes para obtener información sobre las actividades que actualmente se realizan en la LC. El sitio de la ACL tiene una serie de enlaces a grupos especializados en una u otra rama de esta disciplina.

REFERENCIAS BIBLIOGRÁFICAS

Aho, Alfred; Hopcroft, John; Ullman, Jeffrey. 1983. *Data Structures and Algorithms*. MA. Addison Wesley Publishing Company. Reading,

Bratko, Ivan. 2000. *Programming Prolog for Artificial Intelligence*. 3rd edition. Wockingham Addison-Wesley, Longman.

Chomsky, Noam. 2000. *New Horizons in the Study of Language and Mind*. Cambridge. Cambridge. University Press.

Cole, Ronald A. 1997. *Survey of the State of the Art in Human Language* Technology. Cambridge. Cambridge University Press.

Fillmore, Charles. 1968. The Case for Case. In: Bach, E. W. and Harms, R. T. (Eds.). *Universals in Linguistics Theory*. New York. Holt, Rinehart & Winston. pp. 1-88.

Gazdar & Mellish. 1989. *Natural Language Processing in Lisp/Prolog*. Wokingham. Addison Wesley.

Grune Dick, Jacobs, Ceriel J H. 1990. *Parsing Technique: A Practical Guide*. Chichester, Ellis Horwood Limited. También disponible en: http://www.cs.vu.nl/~dick/PTAPG.html.

Halliday, Michael A. K. 1985. An *Introduction to Functional Grammar*. London. Edward Arnold.

Hellwig, Peter. 2000. *A Course in Cooking (Parsing Tutorial)*. http://www.gs.uni-heidelberg.de/~hellwig/tutorial.html

Hopcroft, John & Ullman, John. 2000. *Introduction to Automata Theory, Languages and Computation*. MA. Addison-Wesley. Reading,

Hutchins, W. John & Sommers, Harold L. 1992. *An Introduction to Machine Translation*. London. Academic Press.

Hutchins, W. John. 1986. *Machine Translation: Past, Present, Future*. Chichester. Ellis Horwood.

Jelinek, Frederik. 1999. *Statistical Methods for Speech Recognitio*n. Cambridge. MIT Press.

Jurafsky, Daniel & Martin, James. 2000. *Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. New Jersey. Prentice Hall. Upper Saddle River.

Kay, Martin. 1984. Functional unification grammar: a formalism for machine translation. In *Coling-84*. 10th International Conference on Computational Linguistics. 22nd Annual Meeting of the Association for Computational Linguistics. 1984. New Jersey. Morristown. 75-78.

Manning, Christopher D. & Schütze, Hinrich. 1999. *Foundations of Statistical Natural Language Processing*. MA. MIT Press. Cambridge.

McKeown, Kathy. 1984. *Text Generation*. Cambridge, MA. Cambridge University Press.

Partee, Barbara, H., ter Meulen, Alice & Wall, Robert. 1990. *Mathematical Methods in Linguistics*. Dordrecht.Kluwer.

Pereira, Fernando & Shieber, Stuart. 1987. Prolog and Natural-Language Analysis. *CSLI Lecture Notes*. Vol. 10- Chicago. *Chicago* University Press.

Pinker, Steven. 2000. *The Language Instinct*. 2nd edition. London. Penguin Books.

Pollard, Carl & Sag, Ivan. 1994. *Driven Phrase Structure Grammar*. Chicago. University of Chicago Press.

Pustejovsky, James. 1995. *The Generative Lexicon*. MA, Cambridge. MIT Press.

Sedgewick, Robert. 1990. *Algorithms in* C. MA. Addison-Wesley Publishing Company. Reading.

Shannon, Claude. 1948. A mathematical theory of communication. *Bell System Technical Journal. 27: 379-423, 623-656*. También disponible en: http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html

Shannon,Claude. 1951. Prediction and entropy of printed English. *Bell System Technical Journal*. 30: 50-64.

Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, MA.Addison-Wesley.

ENLACES VARIOS:

Association for Computational Linguistics: http://www.cs.columbia.edu/~acl/

Sistemas de conversión de texto a voz:
http://www.icp.inpg.fr/ICP_old/equipes/synthese/musee.en.html.

Investigación del Instituto Politécnico de Mons, Bélgica, en el área de reconocimiento de voz y conversión de texto a voz. http://tcts.fpms.ac.be