DANIELA RATTI
Istituto per i Circuiti Elettronici
Consiglio Nazionale delle Ricerche
Genova – Italia

Comunicación, información y tratamiento automático del lenguaje[1]

La sociedad en que vivimos se define como "sociedad de la información", ya que se caracteriza por el desarrollo y la gran cantidad de información a disposición y por la capacidad de transmisión de la información que supera las barreras geográficas y políticas. La cantidad de información crece constantemente y paralelamente aumenta la necesidad de instrumentos que permitan conocer esta información del modo más rápido y más económico. Una gran cantidad de datos y potentes conexiones telemáticas no ofrecen en sí ningún servicio significativo y utilizable, si no existen los instrumentos que permitan tratar la información y hacerla accesible al mayor número de usuarios posibles.

Los sistemas informáticos del mañana harán siempre más uso del lenguaje natural y, para permitir un acceso democrático a la información por parte de cualquier ciudadano del mundo, se tendrá que tomar en consideración y encontrar soluciones a los problemas múltiples que la misma información va presentando.

La creación de estos sistemas, por lo tanto, tiene necesidad de mucha elaboración lingüística que permita la *interface* en lenguaje natural (parser morfológicos y sintácticos, interpretación semántica, diccionarios monolingües), sea la elaboración lingüística de los textos o de cualquiera otra fuente informativa (*indización* de las palabras sobre la base de sus significados), sea la planificación y la creación de diccionarios conceptuales multilingües que, utilizados para la interpretación de las interrogantes y para los índices semánticos de los textos, permitan la traducción de expresiones lingüísticas, con los conceptos adecuados, entre una lengua y otra.

Actualmente, de hecho, el panorama mundial está caracterizado por la necesidad de componentes lingüísticos fuertes, capaces de tratar usos reales de la lengua en distintos contextos situacionales, y los que trabajan en el sector se están orientando a reconocer que el desarrollo de sistemas eficaces que responden a necesidades reales, depende de la disponibilidad de componentes con características que permitan tratar la variedad de los fenómenos lingüísticos que se presentan en el uso real de las lenguas.

Las líneas guías que individualizan la investigación y el trabajo de estos años son:

- **construcciones de corpus textuales**
- **construcciones de diccionarios y léxicos**
- **creación de instrumentos y paquetes software para la gestión, el acceso, la custodia y la posibilidad de hallar la información en distintos sectores**

La construcción de corpus textuales de lengua escrita y hablada, seleccionados de manera que puedan representar la variedad de la lengua, con registros oportunos para las

[1]Conferencia Magistral en este Simposio.

distintas unidades a los niveles lingüísticos (morfológicos, sintácticos, semánticos, etc.), es todavía uno de los problemas actuales.

La idea de construir corpus lingüísticos nació entre los años veinte-treinta con el fin de recoger la mayor cantidad posible de material para estudiar el fenómeno de la adquisición del lenguaje y todavía hoy, a pesar de la revolución chomskiana, el corpus queda como un posible método de investigación que puede ser utilizado en todos los sectores de la lingüística: morfología, sintaxis, semántica.

La recopilación de los diccionarios de hoy, sobre todo para las lenguas más desarrolladas desde el punto de vista tecnológico, por ejemplo el inglés, se funda esencialmente sobre corpus lingüísticos y así muchas investigaciones tratan de encontrar soluciones al problema de la homografía, al problema del reconocimiento y producción de los derivados, al fenómeno de los argumentos del verbo y a su régimen, etc. Además, los corpus se utilizan en sociolingüística, en estudios de estilística, en la enseñanza de lengua y lingüística, en estudios de dialectología, en sociología, en psicología, etc.

Los corpus creados en los últimos diez años son muchos, sobre todo para el inglés y, a pesar de ser de la misma lengua, están construidos con muchas características distintas unos de otros.

Aquí merece la pena recordar, a pesar del hecho de que se trate de uno de los corpus más antiguos, el **Corpus de Español de México** realizado por Luis Fernando Lara, que seguramente ha sido el precursor de estas técnicas para la lengua española y, en particular, para la lengua española de Hispanoamérica.

Ese corpus está formado por 2 000 000 de formas y contiene textos escritos y hablados de la mayoría de los estilos lingüísticos y sobre todo está archivado con técnicas estadísticas muy elaboradas que le permiten representar una óptima muestra de la lengua.

Para la lengua española merece también recordarse el **Arthus (Archivo de textos hispánicos de la Universidad de Santiago de Compostela).**

Este Corpus contiene textos pertenecientes a diferentes etapas de la historia del español. La parte contemporánea comprende textos narrativos, teatrales, ensayísticos, periodísticos y orales procedentes de España e Hispanoamérica con un total de 1 450 000 formas. La parte medioeval y clásica contiene un total de 1 063 969 palabras.

Los problemas, objeto de discusión en la creación de corpus, son esencialmente la codificación del material, la organización continua del material y la organización de corpus multilingües.

La codificación de los materiales tiene que ser automatizada lo más posible y lo más extensamente posible, también por eso se necesitan instrumentos capaces de hacerlo. Los identificadores de cada palabra tienen que cubrir más niveles: morfológico, sintáctico, semántico y, además, la palabra tiene que ser dividida en sus componentes, lexema y morfema y no debe aislarse de las otras palabras, es decir, se tiene que tener en cuenta también las relaciones sintácticas entre las palabras. No obstante, la creación de los instrumentos necesarios está todavía atrasada respecto a la creación de corpus y por lo tanto son muy pocos los que contienen por lo menos algunas de estas características, pues todos se diferencian unos de otros.

El proyecto llamado **Amalgam** (**Automatic Mapping among Léxico-gramatical Annotation Models**) constituye un esfuerzo para crear un conjunto de algoritmos capaces de dibujar en mapa los mayores o mejores esquemas de identificadores utilizados en la investigación, de modo que se pueda llegar a la medida estadística de la estructura de la oración en una lengua, objetivo hoy muy difícil, ya que cada corpus tiene una extensión limitada.

Con este proyecto se quiere crear un "Multi-Tagger Corpus", es decir, un corpus que englobe en un único instrumentos el fruto de las comparaciones de los distintos instrumentos.

De hecho la clasificación gramatical usada en cada corpus depende mucho de los aspectos que cada investigador busca mediante su corpus y, por lo tanto, hay corpus con una descripción gramatical muy esmerada y otros con una descripción de nivel mínimo.

Hasta ahora el proyecto ha llegado a producir un "word-taggers" y un "parser" para codificar los textos según los distintos modelos de anotación léxico-gramatical; se ha producido también un prototipo MULTITREEBANK, un corpus de textos en lengua inglesa en el cual cada oración está registrada según distintos modelos de registro léxico-gramatical.

Otro problema es que en la organización del material no se prevé, en la mayoría de los corpus conocidos, un enriquecimiento continuo de la obra y esto produce que los corpus lleguen a ser obsoletos.

En la Universidad de Birmingham el equipo de investigadores de **Cobuild** añade cotidianamente al corpus textos nuevos y este corpus es conocido como **monitor corpus**. Creo que no hace falta repetir a los lingüistas la riqueza que representa un monitor corpus, si se piensa sólo en la posibilidad de averiguar el nacimiento de palabras nuevas, el cambio de significado de palabras existentes, la duración de algunos neologismos, etc.; no obstante, merece la pena observar que no todos los países pueden o quieren invertir en este tipo de investigación.

El problema de un corpus multilingüe nos parece, por el contrario, un problema secundario ya que, o se trata de corpus monolingües organizados de forma que cada uno pueda elegir y comparar las lenguas que quiera, y entonces el problema es la organización y los instrumentos para realizar las comparaciones; o se trata de corpus de textos paralelos que normalmente se realizan en contextos determinados, como por ejemplo el **Multilingual Corpus for International Cooperation (MLCC),** que comprende textos paralelos (los mismos textos en lenguas distintas) tomados de la Comisión Europea, del Parlamento Europeo, de diarios con temas preferentemente económicos en 6 lenguas europeas (inglés, italiano, alemán, español, francés y portigués). Estos tipos de corpus pueden tener una utilidad práctica muy grande, sobre todo en el mismo ambiente de trabajo de donde han salido, pero presentan algunas limitaciones para la investigación por cuanto se trata de un lenguaje sectorial.

La creación de **diccionarios y léxicos** en "machine readable form" es hoy un hecho. Los diccionarios que salen impresos traen normalmente el origen de los corpus de referencia, están tratados con tecnologías digitales y existen, por lo tanto, en "machine readable form".

La casa Editora Zanichelli en Italia ha puesto por primera vez en el mercado un CD-ROM con ocho lenguas y doce diccionarios implicados. La última edición aparece con doce lenguas y dieciocho diccionarios implicados. No sólo algunos diccionarios como Zingarelli y Disc, dos diccionarios italianos muy conocidos y usados a todos los niveles, incluido el nivel escolar, además de su edición en libros ofrecen a los usuarios un CD-ROM analizable desde muchos puntos de vista. Sobre todo el Disc permite conocer muchas informaciones sobre cada palabra, su fecha de aparición, su etimología, etc.

Pero en muchos casos se trata de poco más que versiones electrónicas del libro impreso. Un diccionario electrónico construido no sólo como archivo de datos, sino con la organización y con todas las informaciones útiles para funcionar en los sistemas de "natural language procesing", necesita un esfuerzo intelectual distinto, ya que tiene que ser el conjunto de interrelaciones entre los distintos niveles de descripción lingüística, nivel morfológico, sintácticos, semántico y situacional.

El proyecto **Euro WordNET (University of Amsterdam, ILC-Pisa, Fundación Universidad Empresa, Novell Belgium, University of Sheffield)** parece ser una posible respuesta a este tipo de necesidades, ya que prevé la construcción de una base de datos de palabras en varias lenguas (danés, italiano, español, inglés-americano) accesibles en red con relaciones semánticas para que sea posible encontrar palabras conectadas entre ellas en la misma lengua o en otras lenguas.

Una base de datos así preparada podría ser utilizada en toda una serie de aplicaciones hasta la traducción asistida y podría mejorar la calidad en la búsqueda de las informaciones, ya que los índices de búsqueda de las informaciones en red varían mucho y piden el conocimiento y la competencia en la terminología usada.

El modelo de este proyecto es **WORDNET,** una red de Inglés-Americano disponible gratuitamente. WordNET contiene informaciones sobre los nombres, los adjetivos, los verbos, los adverbios y contiene además unos conjuntos de lexemas llamados "synset" (sinónimos ligados entre ellos) con relaciones semánticas de hiponimia y meronimia entre ellos. EuroWord Net tendría que desarrollar también un lenguaje independiente de conceptos semánticos que puedan enlazar los distintos lenguajes de la red.

Éste parece un proyecto muy ambicioso, ya que si cada "synset" de WORDNET fuera conectado con su oportuna relación al inglés, WORDNET permitirá un enlace no sólo entre la lengua *x* y el inglés sino, a través del inglés, también a las otras lenguas europeas presentes en EUROWORDNET dando las informaciones esenciales para el uso en traducción automática, en "information retrival" y en otros aspectos aplicativos.

Los **instrumentos, los modelos y los paquetes software** para la construcción, la gestión, el acceso, el hallazgo, la custodia de la información en distintos sectores, son el último aspecto sobre el cual vamos a dirigir nuestra atención.

En el pasado los esfuerzos de la lingüística computacional, sobre todo en el mundo de lengua inglesa, se han dirigido hacia el estudio de algoritmos y reglas capaces de analizar o generar secuencias de palabras; se han puesto a punto distintos tipos y estrategias de "parser" y se ha empleado mucho tiempo en evaluar su eficiencia. Se han estudiado fenómenos muy reducidos de la lengua y sobre todo se han examinado frecuentemente ejemplos o subconjuntos muy artificiales.

La orientación actual es muy distinta, y varias son las razones, entre ellas que el rápido desarrollo de la tecnología haya ofrecido la posibilidad de tratar conjuntos siempre mayores de datos, y que se prefiera integrar los métodos basados sobre reglas con métodos empíricos dirigidos a aprovechar conocimientos, también de naturaleza estadística, extraídos de corpus de textos que representen los usos reales de la lengua.

Muchos son los proyectos donde se trabaja actualmente para mejorar el acceso a la información en todos los sentidos; merece recordar el proyecto **UNL (Universal Networking Language)** coordinado por United Nation University Institute of Advanced Studies (Tokío), que se propone enfrentar uno de los problemas más interesantes de Internet, es decir, el acceso a la red para acceder a documentos en varias lenguas permitiendo a cada usuario utilizar su propia lengua.

El proyecto se propone en el 2000 acceder a 13 lenguas (árabe, chino, francés, inglés, ruso, español, japonés, hindú, indonesio, italiano, mongol, portugués, alemán). Si este proyecto tuviera el éxito esperado, sería posible la comunicación entre usuarios de la red en lenguas distintas y así un texto cualquiera se podrá convertir, haciendo referencia a un lenguaje modelo, en la lengua deseada.

Otro proyecto en desarrollo es **CRISTAL (Conceptual Retrieval of Information using a Semantic dictionary for access in three languages)**.

Este proyecto tiene como fin dar facilidad a los usuarios inexpertos que de hecho necesitan una "interface".

Cristal tendrá una "interface" en tres lenguas (inglés, francés, italiano) y ofrecerá la posibilidad de una búsqueda libre utilizando diccionarios semánticos.

Como ya se ha comentado para todas las aplicaciones de la lingüística computacional es necesario tener muchos recursos lingüísticos: corpus, léxicos y gramáticas.

Estos recursos hoy son considerados infraestructuras indispensables para el progreso de la investigación y de las aplicaciones: los léxicos y las gramáticas proveen los sistemas de "natural language procesing" del conocimiento necesario para los distintos fenómenos lingüísticos, mientras que los corpus, hablados y escritos, son fuentes de conocimientos para construir modelos adecuados para el tratamiento de los usos de la lengua.

Puede no estar claro por qué estos recursos faltan todavía en la mayoría de las lenguas después de más de treinta años de historia de la lingüística computacional. Por una parte la tendencia, hasta la mitad de los años ochenta, a privilegiar, como se ha dicho antes, el tratamiento de fenómenos críticos estudiados por las teorías dominantes que llevaba a utilizar pocos ejemplos frecuentemente construidos artificialmente, léxicos de pocas voces, gramáticas con cobertura reducida y, por otra parte, la construcción de recursos lingüísticos amplios tiene un costo muy elevado y requiere el trabajo de equipos muy numerosos.

En los últimos diez años los especialistas han sido empujados a construir recursos adecuados por factores de tipo científico y comercial: el léxico ha tomado un papel más importante en los modelos lingüísticos, se ha comprobado que extendiendo los subconjuntos

lingüísticos, las reglas halladas y comprobadas no funcionaban o pedían tantas y tales modificaciones que el modelo ya no servía, modelos y estrategias no podían desarrollarse más y parecían haber llegado al límite de sus posibilidades y la evolución de las tecnologías empujaba a crear aplicaciones concretas de gran interés industrial y comercial.

Desde siempre la información ha tenido una función muy importante en los distintos aspectos de la vida económica, profesional, cultural y desde siempre ha existido la distinción, que se podría definir de poder, entre quién sabe y quién no sabe. En estos últimos años del siglo, esta distinción se va marcando cada vez más, por un lado los que tendrán la posibilidad de acceder a la información disponible podrán, por lo tanto, conocer, saber, comprar, vender y los que no tengan esa posibilidad, llegarán a ser ciudadanos del mundo de serie B.

Como estudiosos de lingüística no podemos tomar las decisiones políticas para permitir la máxima difusión de los instrumentos que permiten la circulación de la información, pero sí podemos, con nuestros estudios, facilitar la accesibilidad a la información misma porque la lengua es el medio con que se codifica y se difunde la información.

En los próximos años cada país tendrá que desarrollar sus instrumentos lingüísticos adecuados para el tratamiento automático de su lengua con el riesgo, si no lo hace, de perder su sitio en la sociedad global de la información hasta desaparecer con su patrimonio cultural y la humanidad correrá el riesgo de perder su patrimonio mayor que son sus distinciones culturales.