Terminología y lingüística informática

Juan Carlos Merlo

1. Introducción

Los mapas mentales que las sociedades utilizan para visualizar ciertos procesos culturales suelen ser obstáculos insalvables para la comprensión de éstos.

Mucho han tenido que ver en la configuración de estos mapas en la Argentina las ideologías de todo signo y las interpretaciones interesadas. Bastaría con decir que a siete años de la finalización del siglo muchos siguen visualizando la educación primaria con los parámetros con los que la concebía Sarmiento hace 120 años, la secundaria dentro del marco en que la estructuró el Ministro Jorge Eduardo Coll en 1936 y la universitaria según los cánones de la llamada "reforma" de 1918.

Buena parte de la resistencia al cambio que se observa en la sociedad argentina tiene que ver con el desconocimiento y la incomprensión de lo que realmente ocurrió en el mundo a poco de terminada la Segunda Guerra Mundial.

Los exégetas del gran proceso de cambio que ha culminado con lo que unos llaman modernidad y otros era post-industrial o post-industrialismo elaboran discursos laudatorios, mientras otros rechazan la evidencia de un mundo intercomunicado e interdependiente, una verdadera "aldea global" en la que los detractores también son parte del sistema. Pero ni unos ni otros suelen explicar el origen de este cambio, ni las causas que han terminado con la vigencia de la concepción fayolista del trabajo y de la producción.

Me propongo aquí señalar algunos hitos fundamentales que explican el proceso que estamos viviendo.

2. Orígenes de la automatización

Concluida la Segunda Guerra Mundial, el Instituto Tecnológico de Massachusetts (MIT) dio albergue en sus aulas y laboratorios de investigación a muchos de los científicos que debieron emigrar de sus respectivos países. Pero los gestores de los descubrimientos que dieron origen a los grandes cambios tecnológicos de esta mitad del siglo fueron discípulos norteamericanos de los grandes maestros que confluyeron en el MIT. Uno de ellos fue el joven Claude Shannon, autor con Warren Weaver del gran descubrimiento que hizo posible la segmentación de los continuos. Me refiero a la definición de la unidad de información, el bit, en su pequeño opúsculo titulado *The Mathematical Theory of Communication*, publicado por la University of Illinois Press, Urbana, en 1949.

Un año antes, Norbert Wiener había publicado otro pequeño libro también precursor y revolucionario. Lo tituló *Cybernetics or Control and Communication in the Animal and the Machine,* y fue publicado por John Wiley & Sons, Nueva York, 1948.

Por estos mismos años iniciaba sus enseñanzas en el MIT uno de los fundadores de los Círculos Lingüísticos de Moscú y Praga, Roman Jakobson. El maestro ruso, creador del movimiento formalista en su país, debió abandonar Rusia en 1936 por la persecución stalinista. El percibió rápidamente el valor que tendrían para las concepciones sobre el lenguaje y el estudio de las lenguas los descubrimientos de Shannon y Weaver, y rescató del olvido la figura del filósofo norteamericano Charles Sanders Pierce, que había fundamentado en los últimos años del siglo anterior los principios de la semiótica o ciencia general de los signos.

No lejos de allí, en la Universidad de Pennsylvania, el joven Noam Chomsky elaboraba sus primeros estudios de lingüística generativo-transformacional en su tesis doctoral de 1955 titulada *Transformational analysis*, y desarrollaba en el MIT su tratado *The logical structure of linguistic theory*.

La metáfora "el lenguaje de las máquinas" había nacido con las primeras computadoras o "procesadoras electrónicas de datos", autómatas capaces de "calcular, comparar y copiar" cualquier secuencia numérica escrita en el sistema binario. Luego se idearon varios códigos de correspondencias entre números y signos gráficos y alfabéticos lo que hizo posible el procesamiento de los signos lingüísticos, de modo que la metáfora adquirió pleno sentido. Por su parte, las nuevas teorías de la sintaxis y la semántica permitieron la creación de lenguajes artificiales, formados por secuencias de caracteres o "palabras", a las que se asignaban "significados", y secuencias de "palabras" o "frases", a las que se dotaban de "sentidos".

Así nacieron los lenguajes "de máquina" o "de bajo nivel" y los "de programación" o "de alto nivel": FORTRAN en 1956, ALGOL en 1958, COBOL en 1960, BASIC en 1965 y luego los más recientes.

La nueva ciencia del *data processing* se bautizó en Francia en 1962 con el nombre de "informática", y la casi totalidad de los términos que utilizaba eran metáforas de términos lingüísticos: palabra, frase, diccionario, código, coherencia, lectura, escritura, carácter, mensaje, y tantas otras.

3. El desarrollo de la terminología como ciencia

El procesamiento de las palabras de los lenguajes naturales fue una de las primeras utopías que se plantearon los lingüistas ante la capacidad creciente de las computadoras. Estas exigían solamente que las "palabras" que debían procesar tuvieran asignado un solo significado, esto es, que fueran unívocas.

Esta exigencia era "contra natura", ya que, por definición, las palabras de todo lenguaje humano están sometidas a los procesos de cambio semántico y, por consiguiente, tienen grados diversos de ambigüedad.

Empero, la necesidad de las ciencias exactas y naturales de contar con términos unívocos para designar conceptos había dado origen, hacia mediados de los años cuarentas, a la formación de una rama de la lexicología, la *terminología*, que estudiaba los vocabularios de las distintas ciencias como sistemas estructurados que se corresponden con los sistemas conceptuales de cada una de éstas.

La nueva ciencia tenía antecedentes valiosos en las nomenclaturas anatómicas elaboradas en Basilea y Jena en la década del treinta. Pero la primera formulación teórica integral fue la que le dio E. Wuster en su obra de 1959 *Introducción a la ciencia de la terminología general y lexicología terminológica*.

En la concepción germánica, la terminología era la necesaria sistematización

de los conjuntos de términos que en las diversas lenguas designan o representan los sistemas conceptuales de cada campo del conocimiento.

Una *terminología* consiste, pues, en un conjunto de términos unívocos de una lengua (palabras, grupos de palabras, frases, símbolos) que se estructuran en un sistema. Pero la *terminología* como ciencia implica además el estudio interlingüístico de los diversos conjuntos de términos que en las diferentes lenguas representan a un mismo sistema conceptual. En suma, es una ciencia cuyo ejercicio corresponde, de hecho y de derecho, a los traductores profesionales.

Paralelamente, el constante devenir de las ciencias supone una dinámica de los sistemas conceptuales. Cada nuevo descubrimiento, aplicación o desarrollo científico implica la creación de nuevos términos (los neologismos) y la obsolescencia de otros.

Por lo demás, como son parte de un sistema, los términos *significan*, tanto por su relación con un concepto, como por sus diferencias con otros conceptos. De tal modo, la dinámica terminológica abarca, además, la transformación semántica constante a que está sometida la relación término-concepto. Piénsese, por ejemplo, en los cambios de significado que han tenido en este siglo términos de la física como "átomo", "molécula", "núcleo", y de la biología como "gen", "cromosoma", "genética" y tantos otros.

El procesamiento electrónico de la palabra vino en auxilio de la terminología cuando la tarea manual de creación de glosarios, vocabularios y diccionarios monolingües y bilingües se había tornado inmanejable. Desde la construcción de los primeros diccionarios electrónicos en el Departamento de Aeronáutica de EE.UU. durante la Guerra de Vietnam, la computación se convirtió en el aliado irremplazable para la ciencia terminológica. Grandes empresas multinacionales y algunas universidades y organismos estatales construyeron bancos de datos terminológicos computadorizados, y ofrecieron sus servicios a usuarios de todo el mundo. Tanto es así, que los países de la Comunidad Europea primero, y los asiáticos e hispanoamericanos luego, han llegado a sostener grandes bancos de datos terminológicos enlazados telemáticamente. Tal el caso de *RITerm* y la *Red Iberoamericana de Terminología* organizada con el apoyo de la Unión Latina.

4. La terminología en el procesador de textos

Hoy, como lo saben quienes utilizan eficazmente los sistemas de procesamiento de textos, todas las marcas comerciales reconocidas de estos logicales cuentan con diccionarios monolingües ortográficos y de sinónimos, y las hay que incluyen también diccionarios bilingües.

Lamentablemente, por falta de capacitación adecuada, muchos usuarios de procesadores de textos los utilizan como simples máquinas de escribir, pese a que los procesadores de las marcas comerciales más difundidas incluyen funciones para el *procesamiento de palabras* que tienen muchas aplicaciones en terminología.

Todos contienen utilitarios para la validación ortográfica de las palabras de un texto, y agregan funciones especiales para construir el glosario de un solo documento, de un conjunto de documentos, o de una materia o rama del conocimiento.

En este punto, conviene establecer la diferencia que existe entre el *procesamiento de palabras* y el *procesamiento de textos*.

Como se sabe, en el lenguaje corriente no especializado, *texto* es una realización verbal, escrita o impresa. En la lingüística moderna y en la informática, en cambio, *texto* es toda secuencia de frases u oraciones dotadas de la coherencia global que le otorgan sus macroestructuras semántica y pragmática.

La base cognitiva de las macroestructuras semánticas de los textos está constituida principalmente por la terminología del sistema conceptual de una materia o área del conocimiento. A su vez, las macroestructuras pragmáticas se rigen especialmente por reglas de inferencia que determinan la coherencia gramatical y estilística de un texto, dentro de los usos adecuados en esa misma materia.

Cualquiera sea el tipo de macroestructura que hayamos adoptado para desarrollar un texto (narración, argumentación, información general, informe científico, comunicación empresaria, informe didáctico, definición lexicográfica, etc.), la construcción de ese texto implicará la utilización de una base de conocimientos terminológicos y la posibilidad de crear fórmulas o reglas de coherencia global para esa macroestructura.

Un procesador de textos contiene herramientas capaces de asistir al usuario de uno y otro modo. Como procesador de palabras, crea y convalida la terminología o vocabulario utilizado en un párrafo o en el documento completo. Y, como motor de inferencias, puede generar las variables que exigen la gramática y el estilo propios de la macroestructura adoptada. Tales son los casos de formulaciones habituales en ciertos tipos de texto (por ejemplo: "de mi/nuestra consideración", "tengo/tenemos el agrado de dirigirme/dirigirnos a usted/ustedes", "lo/los/la/las saludo/saludamos muy atentamente", en textos de comunicaciones empresarias; el voseo, el tuteo y otras fórmulas de tratamiento; los modismos del habla de un personaje, las formas expresivas propias de una región o lugar, en textos de macroestructura narrativa).

El procesador de textos está provisto de las herramientas aptas para personalizar la redacción de un texto en todos estos aspectos, desde el control pormenorizado de las concordancias gramaticales, hasta la revisión y estandarización o subversión de la gramática y el estilo.

Análogamente, un sistema experto para la asistencia al médico en su consultorio puede organizar un banco de datos con la historia clínica de los pacientes, emitir un diagnóstico a partir de los signos y síntomas detectados en una consulta, verificar la terapéutica adecuada para cada caso clínico y consultar un listado de medicamentos aplicables. Como en la construcción de un texto, los diagnosticadores clínicos computadorizados se valen de una base de conocimientos médicos y farmacológicos y de un motor de inferencias que genera las correspondencias entre todas las variables posibles a partir de los datos diagnósticos de un paciente.

5. La traducción computadorizada

En la actual etapa del desarrollo de la lingüística informática, la traducción computadorizada ha dejado de ser la utopía de los primeros años del procesamiento de textos. Hasta podemos considerar inoperante la vieja discusión semántica sobre la *traducción automática* o la *traducción asistida por computadora*.

La versión interlingüística de textos es una forma de procesamiento en la que la base de conocimientos está constituida por dos diccionarios: uno para la lengua fuente y otro para la lengua meta, con sus respectivas reglas de aplicación de variables gramaticales y estilísticas.

La traducción automática consta de tres fases: una de *análisis* del texto en la lengua fuente, otra de *transferencia de reglas* de la lengua fuente a reglas de la lengua meta, y una tercera de *síntesis* hacia el texto en la lengua meta.

En la fase de la transferencia de reglas se plantean las dificultades de *ingeniería lingüística*. Si se elige el camino de la transferencia a través de un diccionario bilingüe, en el que cada acepción de cada palabra de la lengua fuente tenga su correspondencia en una acepción de una palabra de la lengua meta, se plantea la cuestión del costo de gestión de una base de datos de tales dimensiones para el léxico corriente, al que hay que agregar el costo de las extensiones de la base por el agregado de una o varias terminologías científico-técnicas.

De allí que se haya visto con interés la ingeniería de la *lengua pivote*, natural o artificial. Una lengua del tipo aglutinante, el aymara, ha sido utilizada como pivote en el sistema ATAMIRI ideado por el matemático boliviano Iván Guzmán de Rojas. En esta ingeniería, todas las lenguas se confrontan con la estructura morfosintáctica superficial de la lengua pivote. De este modo se resuelven la mayoría de las ambigüedades hasta el nivel sintáctico. Pero con esta ingeniería quedan sin resolver, y para una etapa de post-edición, las ambigüedades propias de la polisemia en los niveles semántico y pragmático.

Nuestra propia experiencia en la traducción computadorizada de textos de la bibliografía mundial en ciencias biomédicas demuestra que la transferencia interlingüística en los pares de lenguas inglés-español, español-inglés, inglés-portugués, portugués-inglés, español-portugués y portugués-español utilizando computadoras con procesadores Intel 386 y 486 y una memoria mínima de 400 megabytes es perfectamente posible. Los errores propios del nivel semántico-pragmático han sido resueltos por nosotros mediante la generación de macroestructuras típicas de los informes científicos biomédicos, lo que estamos experimentando con muy bajos índices de error.

Esto ocurre porque cuanto más acotada y unívoca es la terminología utilizada, más exacta será la correspondencia entre las frases de las lenguas involucradas. De allí que la traducción interlingüística automática de macroestructuras de información científica y tecnológica tenga tan escasos errores en el nivel terminológico y tan pocas ambigüedades.

6. Conclusión

El actual nivel de desarrollo de logicales para procesamiento automático de textos tiene un amplio campo de aplicaciones. La *lingüística informática* abarca campos que han alcanzado un alto nivel tecnológico.

La *fonología computadorizada* ha desarrollado logicales muy eficaces para orientar y facilitar la enseñanza de lenguas extranjeras y la reeducación de sordos, sordomudos, disléxicos, dislálicos, ciegos y amblíopes. Una rama de la fonología, la sintetización vocálica, ha producido ya sintetizadores de la voz humana que leen textos previamente capturados por un *scanner* y un lector de caracteres en cualquier sistema de escritura (*optical character recognition* u OCR). No está lejano el día en que, mediante analizadores vocálicos, se podrán crear sistemas casi infalibles de identificación personal por la voz grabada.

La *morfología computadorizada* ha producido ya autómatas bilingües que tienen incorporada toda la gramática de un par de lenguas como un anexo a un diccionario bilingüe. El logical de la marca *Spanish Assistant*, de Random House, funciona muy eficientemente en su versión 3.3 de 1992.

La *morfosintaxis computadorizada* ha producido varios traductores de relativa eficacia, pero que exigen una cuidadosa tarea de post-edición. Desde el inolvidable SYSTRAN, hoy en uso en la Comunidad Europea, hay varios logicales de buen funcionamiento en la traducción de textos de lenguaje común no especializado.

La *semántica* y la *pragmática* computadorizadas se encuentran en avanzado nivel experimental, con resultados exitosos en el análisis y traducción de textos científicos y técnicos, como ya se ha dicho.

Y, para terminar, la *terminología computadorizada* ha sido la rama de la *lingüística informática* que mayores logros ha alcanzado en los distintos países de la comunidad internacional, lo que se demuestra por los congresos y conferencias que se ocupan del tema. En nuestra área lingüística, el 4º Simposio Iberoamericano de Terminología y Asamblea de RITerm se reunirá en Buenos Aires en octubre de 1994.

Tan pronto como la formación y capacitación de los traductores profesionales adquiera el nivel de posgrado que debe tener y se desembarace de su dependencia respecto de los estudios jurídicos, se abrirá para la traducción un insospechado campo de nuevas actividades. El sistema productivo y la comunidad científica nacional necesitan urgentemente de estos nuevos profesionales para que la inserción de la Argentina en el mundo deje de ser una utopía o una mera frase para esgrimir en las tribunas políticas.

Juan Carlos Merlo
Director del Centro de Investigaciones para Industrias del Lenguaje (CIPIL)
Director del Centro Terminológico de la Sociedad Iberoamericana de Información Científica (SIICTERM)
Miembro de Número y Vicepresidente de la Academia Argentina de la Comunicación

Contexto e ironía en la traducción de textos literarios

Miguel Angel Montezanti

En el capítulo IX de *Don Quijote de la Mancha*, el ambiguo narrador que ha contado las peripecias del caballero a partir de la declaración confusa de sus fuentes sorprende con la interrupción del relato en un momento en el que, puestas las armas en alto, don Quijote y un vizcaíno están a punto de descargar "dos furibundos fendientes" capaces de tajar sus cuerpos de arriba abajo.

El narrador desconcertado revela entonces que cierto día, mientras recorría el Alcaná de Toledo, dio con un muchacho que deseaba vender unos papeles a un sedero, y que al ver los caracteres, que eran arábigos, y al hacerlos traducir, y al saber que hablaban de una Dulcinea del Toboso, la mejor saladora de puercos de la Mancha, y al demandar la traducción del principio, que decía *Historia de Don Quijote de la Mancha, escrita por Cide Hamete Benengeli, historiador arábigo*, fue tanta su excitación que compró todos los papeles y luego rogóle a un morisco "me volviese aquellos cartapacios en lengua castellana, sin quitarles ni añadirles nada". El morisco "prometió de traducirlos bien y fielmente y con mucha brevedad", aunque el narrador –tal era su expectativa– lo llevó a vivir a su propia casa