El Software de Traducción Multilingüe ATAMIRI

Iván Guzmán de Rojas

Presentación ante el VII Simposio Ibero-Americano de *Terminologia e Indústrias da Língua*
realizado en la Fundación Calouste Gulbenkian, Lisboa, Portugal,
del 14 al 17 de Noviembre de 2000

El Concejo Europeo (Cannes, junio 1995) ha enfatizado la relevancia de los aspectos lingüísticos y culturales de la Sociedad de la Información en Europa. La Conferencia G7 de Ministros sobre *La Sociedad de la Información y Desarrollo*, ha hecho hincapié en el hecho de que las tecnologías de información tienen un potencial formidable para preservar y explotar la diversidad lingüística y cultural. El Documento de la Discusión (julio 1997) *Viviendo y Trabajando Juntamente en la Sociedad de la Información* señala el rol central que debería jugar el multilingüismo en las comunicaciones digitales de gran ancho de banda y en el mundo del Internet.

El *Estudio de productos de traducción automática y servicios* (septiembre 1996) llevado a cabo por el Consorcio Equipe Ltda. en nombre del Concejo Europeo, en su resumen[1] de datos y hallazgos, entre otros aspectos concluye que:

- Algunas de las mejores tecnologías pueden no encontrarse en los productos comercialmente disponibles
- Solamente unos pocos productos son aptos para manejar la demanda de traducción de la CE.
- Los servicios de traducción mejor establecidos son proporcionados actualmente por los vendedores de productos, pero éstos son rudimentarios.

El *Estudio* contiene una lista selecta con datos de 25 paquetes de software de traducción automática: 5 de EU, 3 de Canadá y Alemania, 2 de Dinamarca, Finlandia, Rusia, y Japón, y uno de Francia, Bélgica, Grecia, España, Ucrania y Bolivia. Es notorio que sistemas famosos como ARIANE (Grenoble) y EUROTRA (CE) que hace una década parecían ser muy prometedores, ahora no fueron ni siquiera mencionados en el *Estudio*.

Comparando las características técnicas del software producido por estos desarrolladores, es sorprendente comprobar que solamente uno de ellos, IGRAL de Bolivia, con su software ATAMIRI (aquí presentado), ha sido capaz de diseñar y desarrollar un traductor por máquina **verdaderamente multilingüe**, es decir, un programa, una base de datos gramatical y lexicográfica, que apoya varios lenguajes capaces de operar, ya sea como idioma fuente o idioma meta, con traducción simultánea de cualquier idioma fuente a los varios idiomas meta del conjunto implementado. Todos los demás desarrolladores de software de traducción intentan atender la demanda multilingüe con programas y diccionarios múltiples desarrollados por pares de idiomas, la mayoría capaces de operar solamente en una dirección; pocos son propiamente bidireccionales.

[1] http://158.169.50.95:10080/langeng/reps/mtsurvey/mtsurvey.html

Otro hallazgo interesante en el *Estudio*, es el hecho de que solamente ATAMIRI es capaz de traducir a alta velocidad (más de un millón de palabras por hora en una máquina de 400 MHz bajo WINDOWS NT), mientras todo el resto han registrado velocidades debajo de las 100,000 palabras por hora corriendo en servidores poderosos. Esto significa que ATAMIRI puede fácilmente traducir un página web de 400 palabras en un segundo, mientras otro traductores necesitarían por lo menos 10 a 45 segundos. Estas magnitudes son críticas si nosotros tomamos en cuenta el tiempo adicional requerido para la transmisión de la página web si el sistema traductor está ubicado en el Web donde los suministradores de la información o en el motor de búsqueda, como sería conveniente por razones de gerencia de las bases de datos terminológicas.

Los usuarios consultados en el *Estudio* indican que los sistemas en producción solamente lo hacen con 14 pares de lenguajes unidireccionales, el inglés es utilizado como idioma fuente en 5 pares, el alemán en 3 pares, el chino, finlandés, francés, italiano, ruso y español en solamente un par cada uno. El japonés no fue considerado en el *Estudio*, ya que el japonés es un idioma de baja prioridad para la CE. Por otra parte, el alcance de idiomas cubiertos en proyectos de traducción automática es sorprendentemente amplio, hay en total 162 pares de idiomas ofrecidos en la fecha del *Estudio* y adicionalmente 62 pares de idiomas están en desarrollo. ¿Por qué la discrepancia de 162 pares contra los 14 realmente usados en ambiente de productividad? ¿Es la calidad de traducción pobre, es la velocidad de todo el proceso demasiado lenta, son los costos muy altos?

El **costo alto del multilingüismo** es un factor bien conocido en la CE, que es causado por los requerimientos diversos de clientes y socios comerciales en un contexto donde un creciente volumen del comercio está siendo llevado a cabo electrónicamente a través de las fronteras lingüísticas, donde la competitividad global se apoya cada vez más sobre una mayor disponibilidad de información y eficacia de comunicación.

Los costos estimados para el desarrollo e implementación de N lenguajes en sistemas de traducción de transferencia por pares de lenguajes es proporcional a las **N(N-1)** direcciones del conjunto multilingüe. Mientras que para un sistema basado en la utilización de un lenguaje formal que hace de interlingüa, como es la tecnología de ATAMIRI, los costos son proporcionales simplemente al número[2] de lenguajes **N.**

[2] Si suponemos que el factor de proporcionalidad **k** es en ambos casos aproximadamente igual, la relación R de los costos en el modelo por pares comparado con el modelo de interlingüa es: **R = kN(N-1)/kN = N-1**
Esta ventaja en costos que trae el modelo de interlingüa crece desde R=1 para N=2 (un par de lenguajes) hasta el valor R=20 para el caso de 21 lenguajes, cuya implementación es urgente en el Internet. La experiencia con el modelo de pares de lenguajes, ha mostrado que el factor **k** es del orden de un millón de US$ por lenguaje. Es decir, la comparación para N=21 significa que la ventaja económica del modelo de interlingüa es de 20 veces, es decir, por lo menos 400 millones de US$. Esta gran ventaja crece astronómicamente a medida que aumenta el número de lenguajes a implementarse en ambiente multilingüe, por ejemplo para 101 lenguajes la ventaja es de un factor de 100 veces!

Un ejemplo ilustrativo de los altos costos del multilingüismo, fue el famoso Proyecto EUROTRA al que la CE asignó un presupuesto total de por lo menos 30 millones de US$, para producir un software de traducción multilingüe para ocho idiomas. En casi diez años, no se sabe de un prototipo que haya alcanzado sus objetivos. El sistema ATAMIRI, en el *Workshop* realizado en la OEA, en Washington, en Marzo de 1985, ha demostrado la factibilidad de operar de modo multilingüe a un costo comparativamente bajo, invertido en el desarrollo del software y la implementación de 10 idiomas[3], aunque a diferentes tamaños de diccionarios y niveles de calidad de traducción.

La prueba incontrovertible de que ATAMIRI está construido utilizando una muy avanzada tecnología de ingeniería del lenguaje es el hecho que sus costos de veinte años de R&D han sido cubiertos por el presupuesto personal restringido de su creador, más algunos ingresos provenientes de pocos usuarios y por servicios de traducción a clientes. Por eso es que su diccionario tiene un número relativamente bajo de entradas por idioma, comparado con los productos en el mercado.

La mayor ventaja, económicamente significativa, que ofrece ATAMIRI, es que gracias a su motor de traducción multilingüe conducido por las tablas de su representación sintáctica matricial[4], permite la implementación de un nuevo lenguaje casi sin un esfuerzo adicional de programación. Tan pronto como el nuevo idioma tiene suficientes entradas de léxico en un dominio dado, ya está listo para ser utilizado tanto como idioma fuente o como idioma meta en relación a todos los otros lenguajes ya introducidos en el sistema.

Con ATAMIRI una implantación completa de 16 idiomas, equivalente a 240 direcciones de traducción, tiene un costo estimado en R&D de 10 millones de US$ adicionales[5] a la inversión ya efectuada, mayormente para el trabajo de enriquecimiento terminológico y mejoramiento de la calidad de traducción.

El tiempo de desarrollo y costos requeridos para los sistemas de traducción basados en pares de idiomas conspiran contra la cobertura de más idiomas tan urgentemente requeridos, como aquellos menos hablados en Europa y los idiomas globalmente estratégicos. De esta manera es prácticamente imposible pensar en una red mundial verdaderamente pluricultural.

---

[3] Para el Inglés y el Español, que pueden actuar tanto como lenguajes fuente como meta, la base de datos lexicográfica de ATAMIRI cuenta con algo mas de 25,000 entradas. Con menos entradas a nivel experimental y sólo como lenguajes meta, el sistema traduce a los idiomas: Francés, Portugués, Italiano, Alemán, Holandés, Sueco, Húngaro, y Aymara. La capacidad de implementar mas idiomas es ilimitada.

[4] **ATAMIRI - Sistema de traducción interlingüe utilizando el lenguaje Aymara**
Iván Guzmán de Rojas - Traducción del trabajo presentado en Budapest, Agosto de 1988
Bajo el título de *ATAMIRI - Interlingual MT Using the Aymara Language*
Publicado en: Dan Maxwell / Klaus Schubert / A. P. M. Witkam (eds.)
New Directions in Machine Translation Conference Proceedings, Budapest 18/19- 8-1988
Budapest: John von Neumann Society for Computing Sciences / Dordrecht/Providence: Foris Publishers

[5] Contra los 240 millones de US$ que se requerirían utilizando la tecnología de transferencia por pares de lenguajes, en que la representación sintáctica es arborecente, peculiar para cada par.

En traducción de página web completa, aun el traductor humano confronta la tarea tediosa y difícil de discriminar entre términos que apropiadamente pertenecen al ente del texto traducido y aquellos que se intercalan como marcadores del lenguaje del hipertexto. Los sistemas traductores convencionales a menudo confunden aquellos marcadores que sintácticamente pertenecen a la estructura de la oración, y por tanto deben ser relocalizados de acuerdo a la sintaxis del lenguaje meta, empero los descolocan de modo que los enlaces o términos subrayados no son los correctos en la traducción. Este tipo de transformaciones sintácticas mixtas son manejadas muy bien por ATAMIRI gracias a la generalidad de la representación matricial de lenguaje, subyacente en el diseño de su motor de traducción.

El desarrollo y promoción de guías y normas de interoperabilidad para base de datos de lenguaje y sus componentes se hacen casi imposibles de lograr, a menos que una genuina tecnología multilingüe sea aplicada. Esta tecnología parece hasta ahora encontrarse funcionando desde el año 1985 en el sistema ATAMIRI solamente, aunque sólo como un prototipo de sistema apenas explotado en todo su potencial por falta de apoyo económico para su mayor desarrollo e implementación a nivel operacional.

Se requiere una amplia cooperación mundial para movilizar los talentos necesarios para confrontar creativamente la difícil cuestión de multilingüismo. Como el creador de ATAMIRI, apelo a los líderes de las instituciones y corporaciones que promueven proyectos de Ingeniería del Lenguaje y autoridades de gobierno preocupadas por la problemática de Tecnologías del Lenguaje Humano, para brindar su apoyo a las actividades necesarias para obtener una completa evaluación del sistema ATAMIRI y para probar su tecnología multilingüe, y también medir su capacidad de mejora de la calidad de traducción y velocidad operacional, especialmente en traducción de hipertextos en el Internet.

Iván Guzmán de Rojas
IguzmanRR@hotmail.com