Un aporte tecnológico para resolver el problema de multilinguismo en el Internet

Iván Guzmán de Rojas
igr@atamiri.cc
Consultor en Informática, Investigador en Ingeniería del Lenguaje
miembro de número de la Academia Nacional de Ciencias de Bolivia

Ponencia en el taller sobre *La normalización lingüística y las industrias de la lengua*, realizado durante el 2do. Seminario Interamericano sobre la Gestión de las Lenguas,
Asunción, Paraguay (4 al 6 de junio de 2003)

Un nuevo ámbito de comunicación plurilingüe

Este año celebramos la primera década de operación de esa maravillosa red mundial de comunicación: *Internet*, o también designada como *la web*, en la avasallante jerga del *ciberespacio*. Sin duda, en ese corto tiempo, el Internet se ha convertido en el ámbito más grande de comunicación multilingüe. Sin embargo, está aún lejos de constituir un espacio genuinamente plurilingüe, en el que no solamente estén presentes varias lenguas del mundo, sino que también "se comuniquen" entre ellas.

Cuando nació el Internet, su primer uso fue el fácil acceso a la información tecnológica. Actualmente el c*iberespacio* ya no es solamente un inmenso *réservoir* de toda clase de información, instantáneamente accesible, es además un lugar de encuentro, en el que la comunicación inmediata entre personas o grupos de interés común adquiere cada vez mayor preponderancia.

Los servicios de comunicación disponibles en la red, desde el tan utilizado correo electrónico, hasta los salones de *chat*, los mensajeros, las comunidades virtuales o foros y las reuniones de trabajo virtual, constituyen formidables oportunidades para las relaciones humanas, ya sea como un mero entretenimiento o como un intercambio de opiniones de carácter profesional o también político. El sistema educativo y las formas de plasmar una democracia participativa tienen en estos servicios un formidable desafío para aprovechar las ventajas de un mundo globalizado evitando ser víctimas de sus secuelas aplastantes.

Actualmente, en el directorio del popular buscador *Google*, ver:

http://directory.google.com/Top/World/

a la fecha, fuera del inglés, están registrados 70 idiomas con más el dato de sus respectivos números de *páginas web* en que se leen. Entre ellos, sobresalen, de lejos, el alemán (343,832), español (136,581), francés (125,056) e italiano (115,244) con más de 100,000 *páginas web.*

En el grupo intermedio, con una presencia de más de 10,000 páginas, se encuentran los siguientes 14 idiomas: catalán (27,958), chino (14,508), checo (10,000), danés (35,209), japonés (49,485), coreano (12,187), holandés (59,607), noruego (14,465), polaco (83,162), portugués (12,632), rumano (10,000), ruso (20,301), sueco (44,182) y turco (14,270).

Si bien es cierto que el incremento de la presencia de las lenguas del mundo en el Internet es impresionante, sobre todo si se considera el rol predominante que juega el inglés, por otro lado no debemos olvidar que en el planeta se hablan aproximadamente cinco mil idiomas. Es poco plausible que en los próximos diez años, el número de lenguas con contenidos publicados en el Internet sobrepase el centenar.

También debemos aclarar que la importancia de los idiomas presentes en el Internet no se mide por su número de hablantes, sino por el número de *páginas web* que tienen contenidos en esa lengua. Así se explica que en el directorio de *Google* no figuren lenguas que tienen millones de hablantes en más de un país, como por ejemplo el aymara, el qhechwa o el guaraní, que actualmente no tienen presencia en Internet. En cambio, lenguas como el eusquera, con menos de un millón de hablantes, gracias al apoyo institucional que recibe, ya tiene casi 5,000 *páginas web*. Lo que cuenta en el Internet es la generación de contenidos por parte de los hablantes de una lengua.

La gestión de las lenguas en la perspectiva plurilingüe

El multilingüismo en la red se caracteriza actualmente por islas lingüísticas aisladas unas de otras, con muy pocas posibilidades de intercomunicación entre ellas. En una perspectiva genuinamente plurilingüe, quizás en una utopía a la que deberíamos procurar acercarnos, un hablante de alguna de las lenguas debe poder comunicarse con cualquier hablante de alguna de las otras lenguas. La misma necesidad existe en cuanto al acceso de *páginas web;* dicho de modo simple, todos deberíamos poder leer todo lo publicado en cualquier idioma.

En un ciberespacio con N lenguas presentes se hacen necesarias N(N-1) direcciones de traducción; es decir, ¡ahora para N=70 ya se requieren atender 4,830 direcciones de traducción! Por el modo interactivo en que se navega en el *web* y por la manera instantánea de intercomunicación en los servicios de mensajes, *chat*, foros y comunidades virtuales, cuando decimos "traducción", inevitablemente nos referimos a la *traducción automática*, con todas las deficiencias de las que esta tecnología todavía adolece.

Los sistemas de traducción que ofrecen servicios en el *web*, son de carácter comercial, y apenas cubren una pequeña fracción de esas 4,830 direcciones de traducción requeridas. El sistema que ofrece servicios con más pares de programas de traducción es el clásico SYSTRAN, creado por Peter Thoma en el año 1954 para el par EN <-> RU.

Las direcciones de traducción que actualmente ofrece SYSTRAN son:

EN -> CHs, CHt, DA, NL, FI, FR, DE, GR, IT, JA, KO, NO, PT, RU, ES, SV	16
EN <- CHs, CHt, NL, FR, DE, GR, IT, JA, KO, PO, PT, RU, ES	13
FR -> NL, DE, GR, IT, PT, ES	6
FR <- NL, DE, GR, IT, PT, ES	6

Los programas de este sistema traductor sólo atienden 41 de las 240 direcciones de traducción posibles con ese juego de 16 idiomas. En 29 pares el inglés actúa ya sea como lenguaje fuente, o como meta. En los restantes pares es el francés el que así actúa. No se ofrece la traducción en *direcciones transversales*, como por ejemplo: PT <-> ES.

Es comprensible que las empresas desarrolladoras de sistemas de traducción automática hayan priorizado los pares de idiomas mas interesantes desde el punto de vista del mercado de traducción. Los costos de desarrollo de estos sistemas son enormes, sobrepasan los dos millones de dólares por dirección de traducción. ¿Qué esperanza habría de obtener el retorno a la inversión para el desarrollo de 4,830 programas y juegos de diccionarios?

Por otro lado, las inversiones en I&D para traducción automática son de alto riesgo, muchos proyectos multimillonarios han fracasado en el intento sin haber logrado alcanzar sus metas. El caso más dramático ha sido el sonado proyecto EUROTRA de la Unión Europea que en la década de los 80, en un esfuerzo multinacional, ha insumido más de 50 millones de Euros.

Estos hechos de la realidad del multilingüismo en el Internet nos muestran que la tecnología de traducción automática por pares de idiomas conspira contra el proyecto plurilingüe, ya que es excluyente de las lenguas minoritarias, sin que esta haya sido necesariamente la intención, y además, privilegia un lenguaje dominante.

La meta de una red de *intercomunicación* mundial plurilingüe genera nuevas necesidades instrumentales que plantean un gran desafío a las industrias de la lengua con dos difíciles exigencias tecnológicas:

- **Ingeniería del lenguaje genuinamente multilingüe** capaz de ofrecer servicios de traducción en todas las direcciones que se requieran a costos razonables, proporcionales al número N de idiomas implantados en lugar de la actual proporcionalidad al número N(N-1) de direcciones de traducción.

- **Desarrollo de analizadores y sintetizadores *morfosintácticos*** que permitan manejar bajo un modelo lingüístico universal todas las variantes estructurales de los idiomas con propiedades aglutinantes, y no solamente las gramáticas que siguen el modelo de las lenguas europeas más utilizadas.

Además, para los defensores y promotores de las lenguas "minoritarias", hoy excluidas del Internet, se hace indispensable una tercera exigencia de carácter lingüístico, no tan fácil de cumplir:

- **Desarrollo de léxico y terminología** equivalente para atender los requerimientos de traducción desde y hacia aquellos idiomas que ofrecen los mayores contenidos en el Internet y que ejercen una actividad de comunicación importante.

En nuestro mundo actual, tan intensamente intercomunicado, son muy poco útiles las declaraciones de "idiomas oficiales" de un estado u organismo regional, si al mismo tiempo estas

lenguas permanecen desprovistas de los instrumentos necesarios para hacerse presentes en el Internet, con niveles mínimos de equivalencia lexicológica respecto a los idiomas "importantes".

Aquí me atrevo a pronosticar que aquellas lenguas que en la próxima década no logren afianzarse en la red mundial de comunicación, entrarán en una acelerada e implacable espiral de extinción. La lógica de esta afirmación es sencilla: los jóvenes, que son los portadores de su lengua hacia el futuro, cada vez irán fortaleciendo más su capacidad de comunicación, especialmente la escrita, en el ciberespacio, en sus actividades de formación profesional, comunicación con amigos y grupos de interés, y hasta en entretenimiento. Si no pueden realizar estas actividades en su idioma, lo harán en otro idioma que les sea útil. Su mundo interesante ya no será el de su propia lengua.

El aporte tecnológico del sistema multilingüe ATAMIRI

En marzo de 1985, invitado por la Organización de Estados Americanos, presenté en Washington el primer prototipo del sistema traductor multilingüe, llamado ATAMIRI, que venía desarrollando en La Paz, desde comienzos de los años 80. Después de esa presentación y con esa versión inicial del sistema comenzamos una operación de traducción de documentos técnicos del inglés al español en la Comisión del Canal de Panamá. A pesar de los resultados exitosos, la operación tuvo que ser interrumpida en 1988 debido a que los equipos VS Wang que utilizábamos fueron discontinuados. Desde entonces aprendimos las duras lecciones de tener que ir migrando el sistema a diferentes plataformas operativas a medida que estas evolucionaban al ritmo tormentoso de la década de los 90. Los interesados en conocer la historia de ATAMIRI y los conceptos de su diseño pueden visitar:

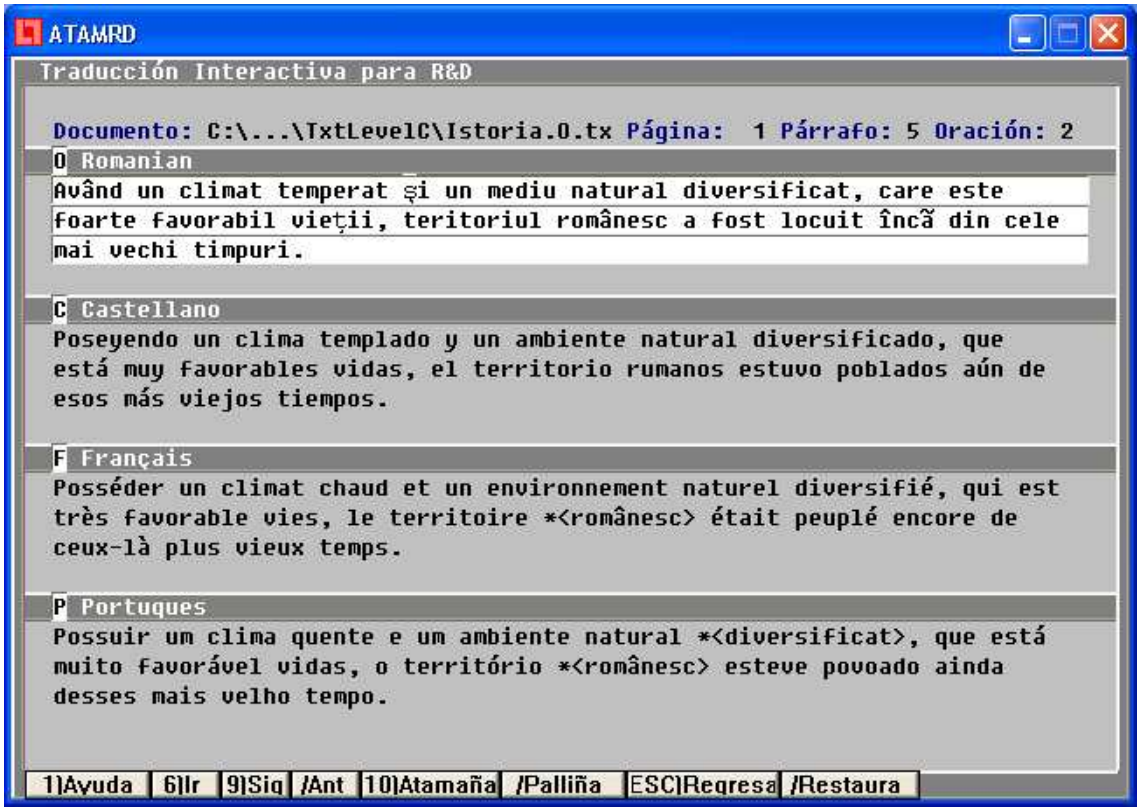
www.atamiri.cc

El año 2001 en Paris, en la UNESCO, presentamos los resultados de la operación piloto de implantación del idioma rumano en el sistema ATAMIRI, experimento que llevamos a cabo en La Paz con nuestro pequeño grupo de Ingeniería del Lenguaje, compuesto por Gladys Dávalos Arze y Marcel Guzmán de Rojas, bajo los auspicios y apoyo técnico de la Unión Latina de París y de la empresa NEOTEC de La Paz, con la colaboración de la Academia de Ciencias de Rumania. Con esta operación piloto quedó demostrada la capacidad multilingüe del sistema que hizo factible que con la introducción de diccionario y reglas gramaticales del rumano, inmediatamente quedaba habilitada la funcionalidad de traducción del rumano desde y hacia los otros idiomas ya implantados anteriormente en el sistema. Las pruebas se realizaron en base a un léxico elemental de 3,200 entradas, y con las tablas de conjugación y declinación del rumano, así como con las reglas sintácticas más frecuentemente utilizadas.

Al año siguiente, también en Paris y bajo los mismos auspicios, presentamos en Internet el servicio de mensajero QOPUCHAWI, con traducción simultánea de los mensajes, en las 30 direcciones de traducción activas para el inglés y los cinco idiomas latinos: ES, FR, PT, IT y RO. La calidad de las traducciones en algunas direcciones todavía no es satisfactoria, se requieren aún ajustes en los algoritmos gramaticales del sistema, y sobre todo, más léxico, incluyendo fraseología típica de los mensajes por Internet. Sin embargo, el servicio, que es gratis, es utilizado por más de 6,000 usuarios registrados desde más de 50 países. Aproximadamente el 40% del

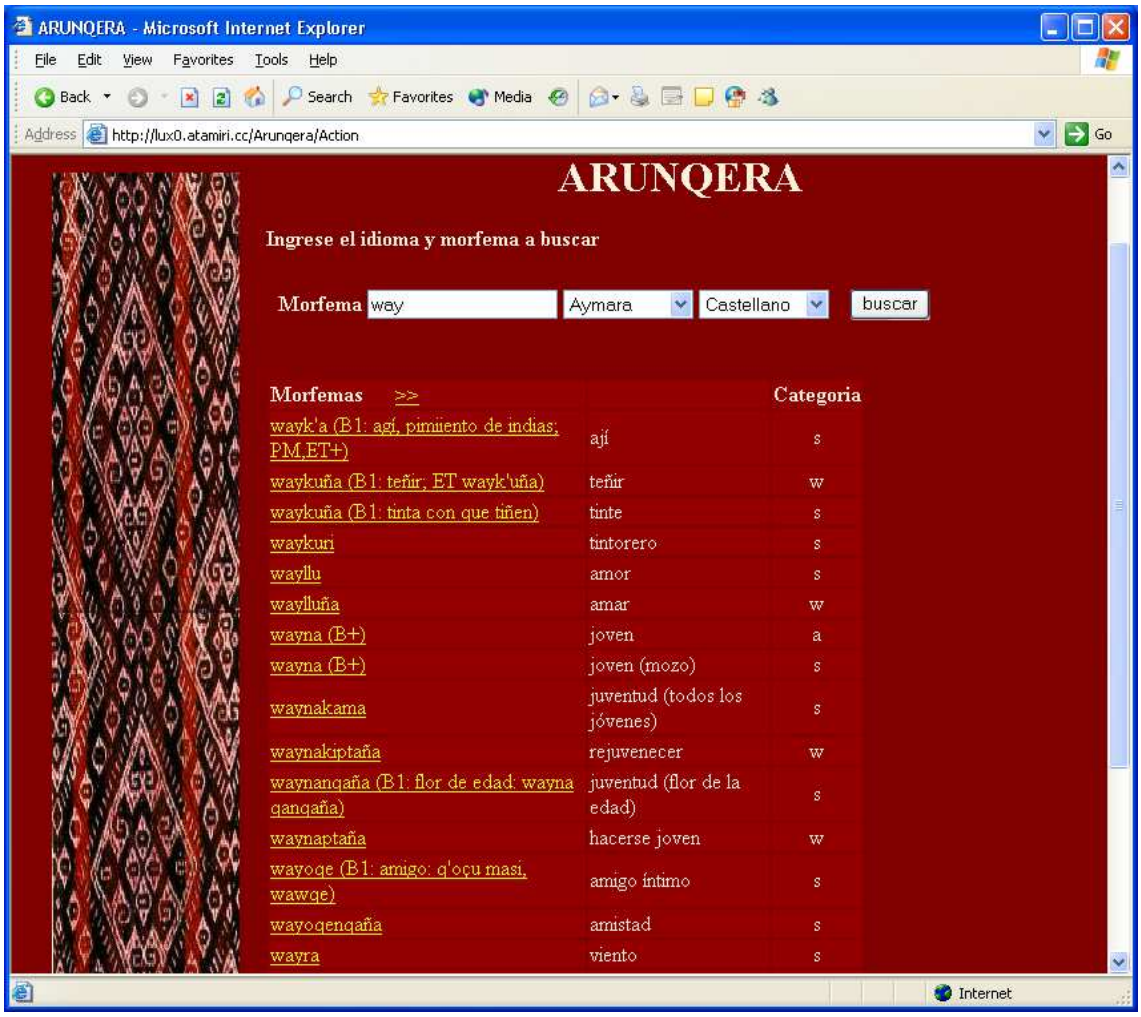
intercambio de mensajes se efectúa en las direcciones transversales, en las que no interviene el inglés. Entre ellas, las más frecuentes son, ES<->FR y ES<->PT.

Para terminar, a continuación muestro dos pantallas del sistema, para ilustrar su utilización con los idiomas latinos:



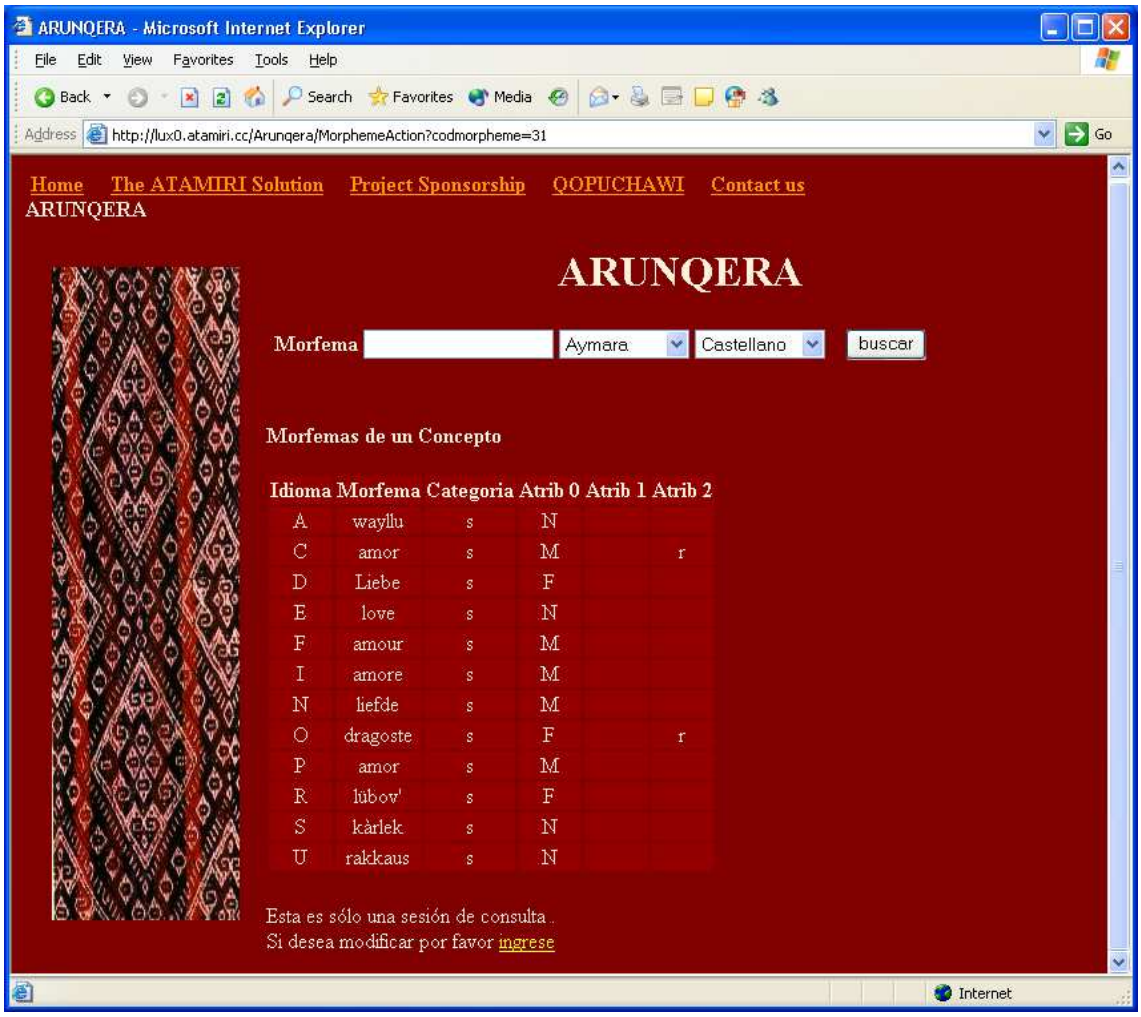
Esta traducción de un trozo de texto en rumano, simultáneamente al castellano, al francés y al portugués, nos muestra cómo el analizador morfosintáctico del sistema desagrega la cadena <teritoriul> en <teritoriu> y el sufijo de articulación <-ul> para efectuar la búsqueda y después el sintetizador en los otros idiomas ordena correctamente el artículo. Si bien las traducciones son aún algo deficientes, son perfectamente inteligibles y más útiles que el texto sin traducir, pese a cierta similitud entre los idiomas latinos.

La siguiente pantalla nos muestra el acceso por Internet a la base de datos lexicográfica ARUNQERA del sistema ATAMIRI, se trata de un módulo que está a prueba para poder consultar e ingresar léxico desde cualquier lugar.



La búsqueda se hizo a partir de la sílaba "way" en aymara, especificando el castellano como idioma para mostrar las equivalencias. Activando el icono >> se obtienen las siguientes 20 entradas de la lista lexicológica.

Se puede activar cualquiera de las entradas lexicológicas del idioma de búsqueda en la columna izquierda, y así se obtienen las equivalencias en los otros idiomas para dicha entrada. Para quienes tengan el registro y contraseña de ingreso, se ofrecen otras pantallas de trabajo que permiten ingresar nuevos términos o efectuar modificaciones a determinadas entradas en el idioma para el que esa persona está autorizada. El sistema lleva una bitácora de los trabajos realizados en la base de datos.



Actualmente ARUNQERA es una pequeña base de datos lexicográfica que le permite realizar al motor de traducción ATAMIRI las pruebas necesarias para el proceso de implantación de idiomas en el sistema. En algunos idiomas se tiene ya suficiente léxico como para efectuar traducciones de buena calidad en ciertos campos temáticos técnicos.

Al 20 de diciembre de 2002, el número de entradas lexicológicas en ARUNQERA, por cada idioma introducido, se agrupaba del siguiente modo:

Número de entradas	Idioma y su código interno
27,205	Español (Castellano C)
21,250	Francés (Française F)
10,509	Portugués (Portuguese P)
12,002	Italiano (Italiano I)
3,204	Rumano (Romanian O)
26,351	Inglés (English E)
15,443	Alemán (Deutsch D)
11,478	Holandés (Nederlands N)

9,816	Ruso (Ruskiy R)
2,643	Sueco (Svenska S)
6,148	Aymara antiguo (A)
1,032	Húngaro (Magyarul M)
166	Turco (Türk T)
18	Finlandés (Suomala U)
67	Japonés (Nippon J)

El último grupo es el de los idiomas en investigación. Los tres últimos idiomas se encuentran en una etapa muy preliminar. Los idiomas con más de 10,000 lexemas ya se prestan para servir como idiomas fuente o meta en traducciones de textos con una exigencia terminológica restringida.

Gracias a la capacidad multilingüe del sistema ATAMIRI, una vez que se haya completado la implantación de cada uno de estos 15 idiomas, el sistema podrá operar en un ambiente multilingüe con 240 direcciones de traducción.

La *implantación* de un idioma en el sistema consiste en:

- Un estudio preliminar de las características lingüísticas del idioma a implantarse y la planificación del trabajo.

- Introducción en ARUNQERA del léxico básico de palabras más frecuentes y la terminología que aparece en los textos a utilizase en las pruebas.

- Introducción en la tabla TUKUNQA los sufijos y prefijos de la morfosintaxis del idioma.

- Pruebas preliminares de conjugación y declinación.

- Introducción de sintagmas en la tabla ARKANAKU.

- Pruebas preliminares de construcción sintáctica.

- Pruebas de traducción al idioma que se implanta, partiendo del inglés y del español como lenguajes fuente.

- Pruebas de traducción del idioma que se implementa, al inglés y el español como lenguajes meta.

- Análisis de la fraseología requerida e introducción de frases.

- Evaluación de la calidad de traducción al y desde el idioma que se implementa, clasificación de anomalías.

- Evaluación de la calidad de traducción utilizando el idioma implementado, como lenguaje fuente y meta en un ambiente multilingüe, con los otros idiomas anteriormente implementados.

- Ajustes de tablas y algoritmos para mejorar la calidad de traducción a un nivel apropiado para prestar servicios.

- Mantenimiento permanente de la base de datos lexicográfica con introducción de terminología, y parámetros complementarios para resolver casos frecuentes de polisemia.

Estas actividades, hasta las pruebas de traducción para verificar la factibilidad de una implantación servible, pueden tomar de tres a seis meses, dependiendo de los recursos humanos con que se cuente. Las demás actividades, hasta que el idioma pueda ser utilizado en ambiente de productividad, a un ritmo de trabajo normal, con recursos razonables, puede tomar entre 18 y 24 meses.

La experiencia ha mostrado que la primera etapa hasta las pruebas de traducción tiene costos del orden de 80,000 Euros, con un nivel lexicográfico de unas 15,000 entradas. De ahí para adelante, los costos dependerán del tamaño y complejidad de la base de datos lexicográfica que se quiera construir. En todo caso, con unos 120,000 Euros adicionales debería ser posible alcanzar un buen nivel de operabilidad en ambiente de producción.

Después de la información aquí proporcionada sobre los logros alcanzados con esta tecnología desarrollada en Bolivia, y el potencial que aún tiene, ruego se me permita hacer un comentario final: considero que ATAMIRI es una tecnología desaprovechada en este mundo en que la problemática del multilinguismo en el Internet se ha hecho tan crítica.