SMART: análisis estadístico multilingüe para la recuperación de información y la traducción

CÁNDIDA VALDÉS / ROBERTO SILVA
Celer Soluciones – Madrid

Introducción

Existe una necesidad creciente e imperiosa de facilitar las comunicaciones entre distintas lenguas. La situación es compleja debido al número de idiomas existentes y a que Internet ha facilitado mucho las comunicaciones entre zonas alejadas. Las publicaciones en la web han aumentado de forma exponencial a medida que aparecían herramientas de publicación dirigidas al público no especializado. Aunque el idioma dominante en Internet es el inglés, el resto avanza muy rápidamente. La presión para disponer de traducciones rápidas entre idiomas muy diversos y facilitar con ello los negocios y las relaciones es enorme.

¿Cómo podemos enfrentarnos a esta situación y a las necesidades que de ella se derivan? Una de las formas evidentes es traducir los textos de un idioma a otro, según sea necesario. Pero esto es costoso y no siempre se encuentran traductores o intérpretes suficientes para todas las combinaciones posibles. Una alternativa válida y frecuentemente utilizada es la de contar con un idioma pivote: los textos, sean del idioma que sean, se traducen a ese idioma común (generalmente inglés, pero en función del caso puede ser otro) y luego se utiliza el documento traducido como base de partida para obtener la traducción final. No obstante, hay muchas situaciones en las que se necesita una traducción inmediata y simplemente es poco práctico o imposible contar con el personal adecuado para esta tarea. Un ejemplo: un particular, por su trabajo, recibe constantemente correos en distintos idiomas, y necesita entenderlos y luego responder a ellos. Un caso distinto es el de un centro de llamadas de un hotel internacional que recibe llamadas desde distintos países, y no siempre es posible encontrar un idioma común para entenderse. Otra situación similar podría ser la de un centro de soporte centralizado para

un producto de software: en este caso las llamadas podrían provenir de cualquier parte del mundo y aunque se podrían derivar a otros centros de soporte locales, lo primero es entender qué está diciendo la persona que llama y proporcionarle una solución. Generalmente, en ninguno de estos casos se puede mantener un grupo de intérpretes suficientemente amplio ni traducir toda la documentación de apoyo a todos los idiomas, o proporcionársela en su idioma nativo en un plazo corto. Hay soluciones alternativas, pero son costosas. SMART[1] pretende ayudar a solucionar algunas de estas situaciones. Inicialmente el proyecto estará limitado a tres combinaciones de idiomas.

Algunas cifras interesantes[2]

Para tener una idea cabal de la importancia que la gestión de la información entre distintos idiomas tiene hoy en día, quizás pueda ser útil tener en mente algunas de las siguientes cifras:

- Número de idiomas vivos: 6 912
- Número de esos idiomas que están al borde de la extinción: 516
- Idioma con el mayor número de hablantes: chino mandarín
- Idioma con el mayor número de hablantes no-nativos: inglés (250-350 millones)
- País con el mayor número de idiomas: Papúa Nueva Guinea (820 idiomas vivos)
- Desde cuándo existen los idiomas: aproximadamente desde 100 000 años aC
- Idioma más antiguo todavía vivo: chino o griego (1 500 aC)
- Idioma con mayor número de palabras: inglés, 250 000 aproximadamente
- Idioma con el alfabeto más grande (incluyendo diacríticos): kabardiano (58 letras). Lo hablan aproximadamente 1 millón de personas que mayoritariamente viven en Rusia.
- Idioma con el alfabeto más corto: rotokas (12 letras). Lo hablan aproximadamente 4 300 personas que viven mayoritariamente en Papúa Nueva Guinea.

---

1 *Statistical Multilingual Analysis for Retrieval and Translation*: Análisis multilingüe estadístico para traducción y búsqueda de información

2 Fuentes: *Ethnologue: Languages of the World* 2005[15], *Oxford English Dictionary*, Wikipedia.org, *Guinness Book of World Records* 2005, UNESCO.

- Idioma con mayor cantidad de sonidos (fonemas): !Xóõ (112 fonemas). Lo hablan aproximadamente 4 200 personas que habitan principalmente en Botswana.
- Idioma con menor número de consonantes: rotokas (6 consonantes)
- Idioma con menor número de vocales: ubyx (2 vocales). Este idioma lo hablan aproximadamente 106 000 personas que mayoritariamente viven en Georgia.
- Idioma con mayor número de vocales: !Xóõ (31 vocales)
- Idioma más publicado: inglés
- Documento traducido a mayor número de idiomas: *Declaración universal de los derechos humanos* (1948), traducida a 321 idiomas y dialectos.
- Las consonantes más comunes: /p/, /t/, /k/, /m/, /n/
- 50% de los idiomas se encuentra en solo 8 países: Papúa Nueva Guinea, Indonesia, Nigeria, India, México, Camerún, Australia y Brasil
- 75% de los idiomas de Brasil (540) han desaparecido desde la colonización portuguesa iniciada en 1530
- 90% de los idiomas de Australia (250) han desaparecido desde la colonización inglesa iniciada en 1800
- 80% de los idiomas del mundo ha desaparecido (esto se basa en una estimación inicial de 30 000 idiomas, pero hay autores que estiman hasta 500 000 idiomas iniciales)
- Hay aproximadamente 300 palabras que son comunes a todos los idiomas (por ejemplo: hombre, mujer, agua, tierra, noche, día).

De los 6 912 idiomas vivos conocidos, 347 (aproximadamente el 5%) tienen al menos un millón de hablantes y en conjunto representan el 94% de la población mundial. En cambio, el 95% de idiomas restantes los habla solo el 6% de la población mundial[3].

Los idiomas e Internet

Internet ha abierto las comunicaciones sin importar las distancias y ha creado un medio en el que los mensajes se transmiten de forma prácticamente instantánea. Esto plantea numerosos retos en cuanto a la organización y comprensión de la información existente. No es fácil separar la información útil de la que no lo es e, incluso cuando se encuentra la información buscada, puede que no esté en un idioma que el

[3] Fuente: *The World Almanac and Book of Facts* 2006 y *Ethnologue: Languages of the World*, <www.ethnologue.com/web.asp>.

usuario entienda. Otro problema adicional es que al realizar la búsqueda en un determinado idioma, el resultado solo encuentre documentos en ese idioma, pero no otros que podrían también ser útiles pero que están en idiomas diferentes.

En términos globales, el estado actual es el siguiente[4]:

Los 10 principales idiomas en Internet					
Idioma	*Usuarios de Internet*	*% del total de usuarios de Internet*	*Estimación de población en 2006, por idioma*	*% que usa Internet del total de hablantes de ese idioma*	*Crecimiento de Internet, por idioma (2000-2005)*
inglés	312 757 646	30,60%	1 125 664 397	27,80%	128,00%
chino	132 301 513	13,00%	1 340 767 863	9,90%	309,60%
japonés	86 300 000	8,50%	128 389 000	67,20%	83,30%
español	80 593 698	7,90%	429 293 261	18,80%	229,20%
alemán	56 853 104	5,60%	95 982 043	59,20%	106,00%
francés	40 974 004	4,00%	381 193 149	10,70%	235,90%
coreano	33 900 000	3,30%	73 945 860	45,80%	78,00%
portugués	32 372 000	3,20%	230 846 275	14,00%	327,30%
italiano	28 870 000	2,80%	59 115 261	48,80%	118,70%
ruso	23 700 000	2,30%	143 682 757	16,50%	664,50%
Total	*828 621 965*	81,00%	4 008 879 867	20,70%	156,00%
Resto de idiomas	194 241 342	19,00%	2 490 817 193	7,80%	421,60%
Total mundial	*1 022 863 307*	100,00%	6 499 697 060	15,70%	183,40%

Estas cifras muestran un dominio abrumador de ciertos idiomas en Internet, en particular del inglés. Más aún, aproximadamente el 90% de los idiomas no se encuentran en Internet y el 80% de los idiomas de África ni siquiera tienen forma escrita (con lo que su presencia en Internet es, en estos momentos, una utopía).

[4] 1) Actualizado a 31 de Marzo de 2006. 2) Es la proporción entre los usuarios de Internet que hablan un idioma y la población total de ese idioma. 3) Una de las fuentes de información más reciente es Nielsen//NetRatings, International Telecommunications Union, *Computer Industry Almanac*. 4) Las estadísticas de población mundial provienen del sitio web de *World Gazetteer*. 5) Las estadísticas se pueden consultar en Internet World Stats.

Como la situación y las necesidades son muy distintas en cada continente, la brecha entre áreas desarrolladas y no desarrolladas en cuanto a uso y presencia en Internet aumenta día a día.

Estadísticas de uso de Internet por zonas[5]						
Zona	*Población (estimación 2006)*	*Población (% del total)*	*Uso de Internet*	*% de la población*	*Uso (% del total mundial)*	*Crecimiento (2000-2005)*
África	915 210 928	14,10%	*23 649 000*	2,60%	2,30%	423,90%
Asia	3 667 774 066	56,40%	*364 270 713*	9,90%	35,60%	218,70%
Europa	807 289 020	12,40%	*291 600 898*	36,10%	28,50%	177,50%
Oriente Medio	190 084 161	2,90%	*18 203 500*	9,60%	1,80%	454,20%
Norteamérica	331 473 276	5,10%	*227 303 680*	68,60%	22,20%	110,30%
Latinoamérica y Caribe	553 908 632	8,50%	*79 962 809*	14,40%	7,80%	342,50%
Oceanía y Australia	33 956 977	0,50%	*17 872 707*	52,60%	1,70%	134,60%
TOTAL	6 499 697 060	100,00%	*1 022 863 307*	15,70%	100,00%	183,40%

Sin embargo, el avance de idiomas distintos del inglés es abrumador. En marzo de 2001 el contenido en inglés en Internet era del 58%. Seis meses después había caído al 51% y en marzo de 2002 se había reducido al 40%. En 2005, la presencia del inglés en Internet es de un 29,7%, lo que representa una reducción del 50% en menos de 3 años.

Recientes estudios indican que un visitante de un sitio web tiene cuatro veces más probabilidades de comprar un producto o servicio si el sitio web se encuentra en su idioma[6] o, al menos, si puede encontrar la página buscada (utilizando un motor de búsqueda como Google) y entender su contenido. Actualmente usan Internet aproximadamente 172 millones de hablantes de inglés y 163 millones que no hablan ese idioma[7]. El 50,4% de los visitantes de páginas web habla un idioma distinto del inglés. Según Forrester Research, un visitante permanece como media el doble en una

5 Datos actualizados a Marzo de 2006. Fuente: <www.internetworldstats.com>.

6 Fuente: <www.idc.com>.

7 Fuente: Global reach <www.glreach.com>

página que se encuentra en su idioma[8]. En la actualidad, un tercio de los sitios web se encuentran en un idioma distinto del inglés[9].

En la Unión Europea, el avance del uso de Internet en los últimos seis años es notorio:

Uso de Internet en la UE[10]			
Año	*Usuarios*	*Población*	*% de la población*
2000	86 796 800	379 634 800	22,90%
2002	143 415 551	379 634 800	37,80%
2004	206 196 749	459 938 780	44,80%
2006	230 396 996	462 371 237	49,80%

Asimismo, la enseñanza de idiomas en Estados Unidos sigue una tendencia parecida. Estas son las cifras de los idiomas con mayor número de estudiantes registrados en Estados Unidos (cifras de 2002[11])

Los 12 idiomas principales que se enseñan en EE.UU.
(basado en registro de estudiantes en otoño de 2002)

Idioma	*%*
español	53%
francés	14,4%
alemán	7,1%
italiano	4,5%
lenguaje de signos	4,3%
japonés	3,7%
chino	2,4%
latín	1,9%
ruso	1,7%
griego	1,5%
hebreo bíblico	1%
árabe	0,7%

8 Fuente: Forrester Research <www.forrester.com>.

9 Fuente: Messaging Online <www.messagingonline.com>.

10 Fuente: <www.internetworldstats.com>.

11 *MLA Newsletter* (primavera de 2004) y *ADFL Bulletin*, vol. 35.2-3 (invierno-primavera 2004).

Métodos de traducción actuales

Las traducciones, hoy en día, se suelen realizar siguiendo alguno de los siguientes métodos[12]:

1. Traducción estándar: se sobrescribe el documento en el idioma original o se crea un nuevo documento en el que se escribe la traducción en el idioma de destino.

2. Traducción automática (MT[13]): sobre un documento electrónico en el idioma original o un soporte físico, se ejecuta una aplicación que traduce el texto al idioma de destino. El resultado depende de la aplicación, del dominio[14] del texto elegido, de la calidad de los diccionarios incluidos en la aplicación, de las personalizaciones que el usuario haya podido realizar en el software para adaptarlo a necesidades específicas, etc. Prácticamente siempre requieren de una etapa final de postedición, a menos que lo que se desee sea simplemente tener una idea básica de un texto en un idioma que no se entiende.

3. Traducción asistida (CAT[15]): el usuario traduce un documento electrónico en el idioma de destino y el resultado se va almacenando, frase a frase, en una memoria de traducción a la vez que se va generando el documento de destino. También se pueden almacenar simultáneamente términos que van conformando un glosario que proporciona ayuda al traductor.

Por último, existen traductores de palabras, que pueden trabajar sobre soportes físicos o electrónicos, pero que no llevan incorporadas reglas gramaticales o estadísticas capaces de analizar una frase. Este tipo de dispositivos equivalen a un diccionario con capacidades de reconocimiento óptico de caracteres.

---

12 La clasificación proporcionada aquí es una visión simplificada de las posibilidades actuales. Frecuentemente las herramientas y los métodos de trabajo se combinan y dentro de cada método hay a su vez subdivisiones.

13 *Machine translation*.

14 La materia sobre la que trata.

15 *Computer Aided Translation*.

¿Qué es SMART?

SMART es un proyecto cuyo fin es desarrollar un programa de traducción automática que proporcione un sistema alternativo en aquellos casos en que sea necesario contar con una traducción inmediata, y no sea factible tener los traductores / intérpretes necesarios, además de proporcionar soporte interno a aplicaciones de búsqueda de información en Internet. Funcionará en ambos sentidos (texto de origen ← → texto de destino) como una aplicación intermedia que permita la rápida comunicación y la solución de problemas entre personas que no tienen un idioma común. Pretendemos que SMART sea capaz de trabajar con soportes electrónicos (documentos, correos electrónicos, etc.) así como también, posiblemente, con voz.

En el consorcio que conforma SMART participan organismos públicos y privados punteros en este tipo de metodologías, así como empresas pioneras en el uso de estas técnicas.

Algunos apuntes sobre traducción automática (MT)

Los programas de traducción automática (MT), para analizar el texto a traducir y generar el texto de destino, se pueden basar en reglas gramaticales, en análisis estadísticos o en una combinación de ambos enfoques.

El análisis mediante reglas gramaticales es el método con mayor número de años de estudio e implantación. Se basa en proporcionarle al programa todas las instrucciones para que sea capaz de analizar gramaticalmente el texto a traducir (formas verbales, combinaciones de palabras, reglas gramaticales, excepciones, listas de abreviaturas, etc.). Esta tarea, que inicialmente parece sencilla, en realidad es muy compleja y tediosa, y generalmente no acaba nunca, ya que el programa tiende a cometer sistemáticamente los mismos errores. Estos errores no siempre se deben a problemas o carencias del programa de traducción automática, sino a la forma en que está escrito el texto a traducir. Una de las posibilidades que existen para evitar este tipo de problemas es controlar la forma en que se escribe el texto a traducir automáticamente. Sin embargo, esto no es una alternativa realista en la mayor parte de los casos.

Para corregir los errores muchas veces es necesario crear nuevas reglas, teniendo cuidado para que no afecten negativamente a las existentes. Esto es especialmente útil cuando los textos a traducir pertenecen a un dominio específico que utiliza un vocabulario y una forma de construir las frases que no sigue los estándares. Por ejemplo, traducir un prospecto

médico no es lo mismo que traducir un guión de cine o un texto comercial. En muchos casos se requiere la intervención de especialistas en lingüística computacional o usuarios avanzados para realizar las modificaciones necesarias en el programa antes de obtener resultados medianamente aceptables. Muchos de los mejores programas comerciales de traducción automática se basan en este tipo de metodología (SYSTRAN, etc.). El problema principal de este tipo de enfoque es el enorme trabajo que continuamente se debe ejercer para que la tarea de postedición[16] no signifique un trabajo mayor a lo que sería haber traducido el texto siguiendo métodos tradicionales. No obstante, en algunas ocasiones no se necesita obtener un texto presentable, sino una idea de lo que dicho texto dice.

El análisis estadístico es el más reciente y posiblemente el más investigado en la actualidad. Una de sus ventajas principales es la rapidez con la que se puede entrenar el motor y obtener resultados medianamente satisfactorios. SMART utilizará un enfoque estadístico aunque, como se verá más adelante, intentará solucionar algunos de los problemas intrínsecos de este tipo de metodologías. Este tipo de metodología se inicio en los 80 y se publicaron los primeros resultados de pruebas en los 90[17]. Consiste en alimentar el motor de traducción con información proveniente de traducciones previas, conocidas y teóricamente correctas. El programa «aprende» en función de la información suministrada automáticamente, en base a parámetros estadísticos que infiere de los textos introducidos. Cuanto mayor sea el número de documentos proporcionados, mejor será (teóricamente) el análisis que de ellos pueda extraer el programa. Luego utilizará este conocimiento para producir nuevas traducciones[18].

En los primeros años se utilizaban pares de palabras, pero pruebas más recientes[19] indican que se obtienen mejores resultados utilizando pares de frases.

---

16 La fase posterior a la traducción automática, consistente en una traducción / revisión del texto proporcionado por el programa de traducción.

17 BROWN *et al.* 1990, BROWN *et al.* 1993.

18 Existe una cierta similitud entre este proceso y el que utilizan los actuales programas de traducción asistida. No obstante, estos programas almacenan en memoria frases traducidas, pero no son capaces de generar una traducción completa de una frase nueva que no coincida con la información presente en la memoria. Solo pueden mostrar lo que ya tienen almacenado y traducir terminología, si es que disponen de un glosario asociado a la memoria.

19 OCH *et al.* 1999, MARCU / WONG 2002, KOEHN *et al.* 2003, OCH / NEY 2004.

Muchos coinciden en que el método estadístico es, hoy en día, mejor en la mayor parte de las situaciones que el método basado en reglas gramaticales. Naturalmente, esto es motivo de debate entre las distintas líneas de investigación. No obstante, ninguno de los dos es perfecto y los avances en ambas direcciones son continuos.

Problemas de los métodos estadísticos

Los principales problemas que presenta el método estadístico y que SMART intentará abordar son:

1. Todos los modelos requieren una toma de decisiones respecto al parámetro y fórmula que se utilizará para calcular la mejor traducción posible de una frase. Sea cual sea la elección, esto influirá directamente sobre la búsqueda que el programa realice y sobre el tiempo que tarde en proporcionar un resultado. Lógicamente, de nada sirve que la traducción proporcionada sea de alta calidad si el cálculo lleva horas. Se necesita, en la mayor parte de los casos, una respuesta instantánea.

2. Los sistemas de traducción automática se basan en un proceso de alimentación que no es simultáneo al de la traducción. En otras palabras, primero se alimentan y entrenan, y luego se traduce. A diferencia de los programas tipo CAT[20], lo que el traductor incorpore solo podrá pasar a formar parte del programa cuando tanto el documento original como la traducción estén finalizados, sean introducidos en el motor de traducción automática y se recalcule en función de esta nueva información. Esto es un problema grave en situaciones reales de traducción en las que es frecuente realizar cambios en función de los requisitos del trabajo o de la identificación de nueva terminología. Además, el traductor puede identificar errores y corregirlos, pero el programa no será capaz de incorporar estas correcciones la siguiente vez que aparezca una frase similar o idéntica a la que acaba de modificar.

3. El enfoque probabilístico depende en gran parte de la cantidad y del tipo de documentación con que se lo alimenta. Esto condiciona los resultados. Si se ha alimentado el motor con gran cantidad de documentación médica, se pueden obtener buenos resultados en nuevos

[20] Programas de traducción asistida, generalmente basados en una memoria de traducción.

textos médicos, pero los resultados serán pobres en textos legales. Un problema grave al que se enfrenta cualquier motor estadístico es la no disponibilidad de grandes corpus bilingües[21] o multilingües en todos los temas e idiomas deseados. Los resultados obtenidos al entrenar el motor con unos pocos documentos son deficientes.

4. Una desventaja de los modelos estadísticos respecto a los modelos de reglas gramaticales es la poca capacidad para producir un texto de destino gramaticalmente sólido. Al basarse en modelos estadísticos simples, se gana en rapidez para entrenarlos, pero se pierde la capacidad de analizar la frase generada y construir una que sea aceptable y siga las reglas gramaticales del idioma de destino. La traducción se realiza palabra a palabra, pero el programa no se preocupa de construir una frase correcta.

Objetivos de SMART

SMART pretende buscar respuestas a algunos de los principales problemas actuales de la traducción automática. Para ello tomará en cuenta lo siguiente:

1. Se seguirán los últimos avances en análisis sintáctico para enfocar el problema desde el punto de vista de problemas de optimización globales de funciones de muchas variables. Se piensa que esta estrategia puede llevar a avances significativos dentro del modelo estadístico.

2. Se intentará crear un modelo adaptativo de traducción automática que responda a los cambios realizados por los traductores respecto al texto propuesto por el programa. Esto es fundamental para obtener una respuesta positiva de quienes se supone que podrían ser los principales destinatarios del supuesto producto comercial resultado de este proyecto.

3. Se estudiarán modelos de adaptación específicos de un dominio, tanto para los casos en que se dispone del conjunto de documentos en los que se basó la fase de entrenamiento del programa, como para aquellos en que el modelo está disponible, pero no los datos.

---

[21] Grupos de documentos en idioma original, con sus correspondientes documentos traducidos.

4. Se investigará la posibilidad de utilización de métodos que no requieren un analizador sintáctico, ya que son muy difíciles de crear, aunque con ellos se obtienen, en general, buenos resultados. Estos métodos pueden incluir mecanismos de comparación de secuencias de símbolos[22] y análisis de la información morfológica de las frases[23].

5. Abordar, desde la perspectiva de los métodos probabilísticos, los problemas propios de las búsquedas de información y recuperación de documentos en idiomas distintos del propio[24]. Un ejemplo sería el de una consulta en Internet formulada en el idioma del usuario, pero cuyo resultado devuelve correctamente documentos en un idioma distinto del de la búsqueda. Este problema es complejo y se analizarán distintas variantes. El enfoque más común actualmente consiste en traducir la consulta realizada por el usuario y luego buscar en los distintos idiomas. Sin embargo, estas búsquedas no están adaptadas al lenguaje propio de cada dominio, por lo que los resultados no suelen ser buenos. Una posibilidad que se analizará en este proyecto es la extracción de vocabulario específico de cada dominio tanto desde corpus paralelos[25] como multilingües[26], para permitir la traducción de unidades de traducción no contiguas consistentes en varias palabras. Este glosario será, idealmente, actualizable con las modificaciones y nuevas entradas propuestas por el usuario.

6. Se estudiará qué variaciones o combinaciones de métodos de análisis latente no dependiente del idioma[27] son las más adecuadas para el acceso a la información en idiomas distintos del original (CLTIA).

---

22 LODHI *et al.* 2002, SAUNDERS *et al.* 2002, CANCEDDA *et al.* 2003.

23 CORTES *et al.* 2004.

24 *Cross-lingual textual information access* (CLTIA): acceso a la información multilingüe en función del contexto.

25 Conjuntos de documentos originales y su correspondiente traducción.

26 Conjunto de documentos originales y traducciones que aunque no son uno traducción del otro, cubren el mismo tema.

27 Método que no pretende identificar la traducción de palabras utilizando un conjunto de documentos, sino las relaciones conceptuales entre ellos.

Evaluación del sistema

Se pretende que SMART cuente con tres escenarios distintos para probar los resultados que se vayan obteniendo a lo largo de los 36 meses de duración del proyecto:

1. CAT[28]: en este primer escenario, el usuario es un traductor profesional. En los sistemas estándar basados en memorias de traducción, el traductor va incorporando frases a una memoria a medida que traduce. Esta no es la situación en sistemas de traducción automática, lo cual tiene un impacto negativo en la productividad y en la paciencia del traductor. Se estudiará el impacto que puede tener en un sistema de traducción automática la incorporación de un módulo capaz de aprender de las correcciones e incorporaciones realizadas por el traductor. Para ello se utilizarán algoritmos adaptativos en línea. Se pretende que SMART tenga un impacto positivo que resulte en un aumento de la productividad de un traductor profesional de alrededor de un 20% respecto a los resultados iniciales antes de utilizar el sistema.

2. Centro de llamadas con consultas de tipo CLIR[29]: en este escenario el usuario es un técnico que responde llamadas telefónicas para solucionar problemas de otros usuarios. El técnico responde a la llamada en el idioma A y a la vez consulta documentación existente en el idioma B. La consulta que realiza para buscar la información que requiere no la realiza en el idioma en que se encuentra esa documentación, sino en su propio idioma, o en cualquier caso, en el idioma en el que ha contestado la llamada. Cuando ya tenga localizada la información, dispondrá de un sistema de traducción automática (tanto a nivel de palabra como de todo el documento) para poder proporcionar una respuesta en el idioma de la persona que ha planteado el problema. La intención del consorcio es lograr que el uso de SMART suponga una reducción de un 15% en el tiempo medio necesario para resolver una consulta telefónica en un centro técnico de llamadas.

---

28 *Computer-Aided Translation*: traducción asistida.

29 *Cross-Language Information Retrieval*: recuperación de información multilingüe.

3. Consultas a la Wikipedia[30]: en este escenario el usuario realiza consultas a la enciclopedia en línea Wikipedia, en el idioma A, encuentra documentos en distintos idiomas y el sistema le proporciona ayuda para acceder a aquellos contenidos escritos en idiomas distintos del propio o del que ha formulado la consulta. Para este escenario la intención es probar que el acceso a documentación en idiomas distintos del propio y desconocidos para el usuario es posible en tiempo real.

Para estos escenarios se podrán utilizar prototipos de sistemas de traducción automática desarrollados por algunos de los socios del consorcio (Xerox, NRC) y programas de libre distribución como LEMUR. Las mejoras en la productividad se medirán utilizando mecanismos automáticos y mediante experimentos en los que participarán Celer, Amebis y un grupo de Xerox.

Participantes en el proyecto

Xerox – Francia, empresa especialista en sistemas de documentación y servicios. Participará a través de su sede en Grenoble, Francia - Xerox Research Centre Europe (XRCE).
Amebis – Eslovenia, empresa eslovena especializada en el desarrollo y producción de software (principalmente herramientas de traducción) y bases de datos.

30 Wikipedia es una enciclopedia libre multilingüe basada en la tecnología wiki. Wikipedia se escribe de forma colaborativa por voluntarios, permitiendo que la gran mayoría de los artículos sean modificados por cualquier persona con acceso mediante un navegador web. El proyecto comenzó el 15 de enero de 2001, fundada por Jimbo Wales y Larry Sanger como complemento de la enciclopedia escrita por expertos Nupedia. Ahora depende de la fundación sin ánimo de lucro Wikimedia Foundation. Wikipedia tiene más de 3 800 000 artículos, incluyendo más de 1 100 000 en su edición en inglés, y a finales de febrero de 2006 alcanzó la cifra de 1 000 000 de usuarios registrados. Desde su concepción, Wikipedia no sólo ha ganado en popularidad[1] sino que su éxito ha propiciado la aparición de proyectos hermanos. Existe, sin embargo, controversia sobre su fiabilidad.[2] En este sentido, la revista científica *Nature* declaró a la Wikipedia en inglés casi tan exacta como la enciclopedia *Britannica* en artículos científicos.[3] [4]
Los medios de comunicación y la comunidad científica citan a Wikipedia, algunas veces de manera crítica, otras ensalzándola por su carácter de libre distribución, mejora constante y multifacética, no sin mencionar su naturaleza multilingüe. A menudo no es citada como fuente única, sino como fuente informativa complementaria.
Wikipedia está editada en más de 229 idiomas[5], de los cuales solamente 137 están activos (mayo de 2006). Catorce ediciones superan los 50 000 artículos: inglés, alemán, francés, japonés, polaco, italiano, sueco, danés, portugués, español, chino, ruso, finés y noruego. La versión en alemán ha sido distribuida en DVD-ROM y se tiene la intención de hacer una versión inglesa en DVD y en papel. Muchas de sus otras ediciones han sido replicadas a través de internet (mediante «espejos») o se han creado enciclopedias derivadas (*forks*) en otros sitios webs (información tomada directamente de la página principal de Wikipedia).

Celer Soluciones – España, empresa especialista en servicios de traducción multilingüe, con sede en Madrid, España, con 19 años de experiencia en el sector.
Jozef Stefan Institute, centro de investigación pura y aplicada en el campo de las ciencias naturales y la tecnología fundado en 1949, con sede en Ljubljana, Eslovenia.
National Research Council Canadá, uno de los principales centros de investigación y desarrollo de Canadá
University of Bristol – Inglaterra, que participará a través del Departamento de Ingeniería Matemática
University of Helsinki – Finlandia, que participará a través del Departamento de Ciencias Informáticas
Universita degli Studi di Milano – Italia, que participará a través del Departamento de Ciencias de la Información
University of Southampton – Inglaterra, que participará a través del grupo de Imagen, Voz y Sistemas inteligentes.