Luces y sombras en los 75 años de traducción automática

Petra Díaz Prieto
Universidad de León
petra.diaz@unileon.es

Introducción

En las últimas dos décadas hemos sido testigos de grandes cambios. Por un lado, un enorme crecimiento de la tecnología de la información (TI) con las consiguientes ventajas de rapidez, impacto visual, facilidad de uso, comodidad, y la relación coste-eficacia. Con la tecnología de la información ha surgido la cultura de la pantalla que tiende a sustituir la cultura de los documentos impresos, con documentos electrónicos que se pueden guardar, transmitir y acceder directamente a través del ordenador (e-mail, bases de datos y demás información almacenada). Estos documentos están disponibles de forma instantánea y se pueden abrir y procesar con mucha más facilidad y rapidez que el material impreso. Internet con su acceso universal a la información y la comunicación instantánea entre los usuarios ha creado una libertad física y geográfica inconcebible en el pasado, pero que necesita de la traducción.

Por otro lado, con el desarrollo del mercado mundial, la industria y el comercio funcionan, más que nunca, a escala internacional, cada vez con mayor libertad y flexibilidad en términos de intercambio de productos y servicios, a lo que se suma la necesidad de cooperación de lo países en muchos otros campos, como en el educativo, económico humanitario y ecológico, etc. Sin embargo, en este mundo cada vez más globalizado, donde no existen barreras tecnológicas comunicativas y en el que la comunicación es la base del progreso, siguen existiendo barreras lingüísticas al mantenerse la creencia general de que las personas tienen derecho a utilizar su propia lengua. La variedad de lenguas existentes supone un obstáculo para la difusión de información e ideas al no existir una *lingua franca*, a pesar de la importancia del

inglés como lengua vehicular para el 20% de la humanidad, y que supone más del 90% de toda la información depositada en Internet. Se necesita, por tanto, encontrar soluciones a los problemas lingüísticos a fin de permitir la libre circulación de información y facilitar las relaciones bilaterales y multilaterales.

La traducción se presenta como un medio idóneo para acceder a cualquier tipo de información con independencia de la lengua en que esté expresada, No obstante, las agencias y organismos dedicados a la traducción se encuentran saturados, en parte, por la ingente cantidad de material que generan los grandes organismos internacionales y en parte para satisfacer la creciente demanda de traducciones de documentación jurídica, manuales de instrucciones, libros de texto de medicina o electrónica, patentes industriales, panfletos publicitarios, reportajes periodísticos, etc. En la actualidad la demanda de traducciones no se puede atender porque no hay suficientes traductores humanos, ya que, la productividad de un ser humano está esencialmente limitada. Las estadísticas varían, pero en general, un traductor humano, en casos muy favorables, puede procesar unas 20 páginas por día, y cuando se trata de textos difíciles un traductor no puede procesar más de 4-6 páginas ó 2.000 palabras por día. La traducción automática se erige como una solución para combatir el cuello de botella que supone la traducción en esta sociedad de la información. La creación de herramientas para la automatización de la traducción se presenta como un importante avance que permite vencer, al menos en parte, los obstáculos para acceder a la información.

La Traducción Automática (TA), también llamada MT (del inglés *Machine Translation*), es un área de la lingüística computacional que investiga el uso de software para traducir texto o habla de un lenguaje natural a otro tanto con o sin ayuda humana. La TA en sentido amplio abarca toda una variedad de sistemas que sólo comparten la utilización del ordenador como instrumento de traducción. En palabras de la Asociación Europea para la Traducción Automática (European Association for Machine Translation) (EAMT) (1997) "Machine Translation (MT) is the application of computers to the task of translating texts from one natural language to another".

Han pasado 75 años desde que se patentaran las primeras máquinas de traducir, aunque, se puede decir que es a partir de la década de 1940, con la aparición de los primeros ordenadores, cuando la traducción automática se convierte en una de las aplicaciones más importantes de la informática. A pesar de no contar con muchos años de vida, la TA ha pasado por periodos en que ha sido más conocida por sus fracasos que por sus éxitos, y por épocas de esplendor que impulsaron numerosas investigaciones con el objetivo de conseguir una traducción totalmente automática con una calidad cercana a la alcanzada por el traductor humano.

En este trabajo, nuestra intención es hacer una revisión de la trayectoria de la TA no sólo en el plano histórico, sino también repasar cual ha sido su evolución y como el primitivo objetivo de una traducción totalmente automática ha dado paso a la investigación de ayudas y herramientas para el profesional o no profesional que explora las posibilidades del ordenador para apoyar las capacidades humanas. Trataremos el contexto en el que se están desarrollando las actuales investigaciones, así como los sistemas más significativos que han existido y existen.

Consideramos que esta revisión no quedaría completa si no mencionáramos la trayectoria de la TA en España.

1. Historia

El anhelo de conseguir artilugios mecánicos que puedan superar las barreras lingüísticas no es nuevo y ha despertado el interés de grandes estudiosos. Abaitua citando a Martin Kay dice:

> El empeño en conseguir máquinas traductoras ha merecido la atención de algunas de las mentes más preclaras de disciplinas como la lingüística, la filosofía, las matemáticas o la informática. La traducción automática ejerce, además, una irresistible atracción sobre un nutrido grupo de incondicionales ¿A qué se debe esta fascinación? (Abaitua 2002).

Para poder contestar a esta pregunta nada mejor que un paseo por la historia de la TA. En el plano teórico, (sin ninguna influencia en los sistemas actuales), su historia se remonta al siglo XVII cuando Leibniz y Descartes, apoyando la creación de una *lengua universal* no ambigua, basada en principios lógicos y símbolos icónicos, que permitiera comunicarse a toda la humanidad, hablaron de la utilización de diccionarios mecánicos basados en códigos numéricos universales para superar las barreras del lenguaje. Sin embargo, no será hasta mediados del siglo veinte cuando se empiece a hacer realidad la traducción automática.

En el 2008 se cumplen 75 años desde que se puso en práctica la traducción automática y comenzara a hacerse realidad ese sueño. En este periodo de tiempo se han realizado grandes y pequeños experimentos, y se han invertido grandes cantidades de dinero tanto procedentes de diversas instituciones como de grandes empresas. Se ha pasado por épocas de gran optimismo, y se ha apuntado perspectivas de remplazar al traductor humano, como lo demuestran las manifestaciones realizadas por el Dr. A Booth en 1959, cuando la TA se encontraba aún en sus inicios:

> Not only does it now appear possible that translations of good quality can be made from both scientific and literary texts, but also that some of the recent developments in machine technology will make it possible to read directly from the printed or typewritten page (Pinchuck 1977: 239).

El optimismo no es tan grande actualmente y no todos los textos se consideran aptos para traducirse automáticamente en un ordenador debido a la enorme complejidad de traducir el lenguaje humano y a las intrínsecas limitaciones de la generación actual de programas de traducción. Las investigaciones actuales se dirigen más hacia los programas de ayuda al traductor *Computer Aided Translation* (CAT), aunque la traducción totalmente automática y con una calidad alta *Fully High Quality Translation* (FHQT) no se descarta.

Tradicionalmente se ha abordado la evolución de la TA por décadas (Slocum 1998; Hutchins 1995a, 1995b, 2000, 2005; Abaitua

2002), sin embargo, en este estudio hemos preferido dividir el desarrollo de la TA en antes y después del informe *Language Processing Advisory Committee* (ALPAC) que supuso un parón en la investigación de la TA.

Creemos que esta visión histórica de la evolución e interés despertado por la TA no quedaría completa si no mencionáramos la investigación llevada a cabo en España.

1.1. Antes del informe ALPAC

Es en 1933 cuando se dan los primeros pasos prácticos. En este año se solicitaron dos patentes para máquinas de traducir, una en Francia y otra en Rusia. Por un lado la máquina del francoarmerio, George Artsouni, era un dispositivo de almacenaje en banda de papel que permitía encontrar el equivalente de cualquier palabra en otra lengua. Por otro lado está la propuesta mucho más interesante del estudiante ruso Petr Smirnov-Troyanski. Al parecer se mostró un prototipo en 1937. Su sistema incluye no sólo un diccionario bilingüe, sino también un esquema para codificar los papeles gramaticales entre lenguas (basado en el esperanto) y un esbozo de cómo análisis y síntesis podían trabajar. Establece tres fases en el proceso de traducción. La primera fase era la edición en la lengua fuente para organizar las palabras dentro de sus formas lógicas y funciones sintácticas; la segunda fase era en la que la máquina traducía estas formas en una lengua meta; y la tercera fase era la post edición en la lengua meta para normalizar su salida. Aunque su sistema era muy rudimentario, y tan sólo se planteaba la segunda fase como un proceso automático, los principios de su máquina siguen vigentes y sus tres fases en el proceso de la traducción pueden asimilarse a las actuales de análisis, transferencia y generación. Sus ideas no fueron conocidas hasta finales de los cincuenta, cuando los ordenadores estaban en auge.

Los primeros desarrollos serios de la TA se producen poco después de la segunda Guerra Mundial, con la aparición del famoso ordenador ENIAC en 1946. La fecha de inicio viene dada por una carta de Marzo de 1947 enviada por Warren Weaver de la Fundación Rockefeller al cibernético Norbert Wiener.

El británico Dr. Andrew Donal Booth sugirió a Warren Weaver la posibilidad de emplear los ordenadores para ayudar al traductor y para la traducción de lenguajes naturales. De vuelta a Birkbeck College (Londres), Booth comenzó a explorar la mecanización de un diccionario bilingüe y a colaborar con Richard H. Richens de Cambridge, que había estado usando tarjetas perforadas para producir traducciones *en borrador* de palabra por palabra de *abstracts* científicos.

Por su lado, el Dr Weaver consideró que la tarea de construir una máquina de traducir estaba relacionada con el proceso de descifrar códigos enemigos. Su hipótesis de partida era creer que bastaba reemplazar las palabras de una frase inicial por equivalentes para obtener una traducción inteligible. Es decir, vio el lenguaje como un código, en el sentido de que era un sistema limitado de signos y que traducir era reemplazar cada signo en el sistema original por un signo equivalente en el nuevo sistema. Fue en Julio de 1949, cuando dio a conocer públicamente la idea de la TA con la publicación de un informe llamado simplemente *Translation* en el que explicaba (a) que la multiplicidad de las lenguas era un serio impedimento para el entendimiento internacional; (b) que el uso de los ordenadores de alta capacidad ayudarían a resolver el problema; y (c) que la traducción de palabra por palabra tenía grandes limitaciones. Anticipó cuatro posibles métodos para abordar la traducción: el uso de técnicas criptográficas, la aplicación de los teoremas de Shannon en teoría de la información y la utilidad de la estadística, así como la posibilidad de aprovechar la lógica subyacente al lenguaje humano y sus aparentes propiedades universales.

Dicho informe fue la publicación más influyente en la primera época de la máquina de traducir y fue también un estímulo para que, en un breve espacio de tiempo, se iniciaran varios proyectos de investigación sobre la TA, primero en algunos centros y universidades de EE.UU., tal como la Universidad de Washington (Seattle), la Universidad de California en los Ángeles y el *Massachusetts Institute of Technology* (MIT) y más tarde en el resto del mundo.

Los primeros experimentos que se hicieron fueron científicos y de ingeniería, no lingüísticos. Se pensaba que era única y exclusivamente un problema de ingeniería. Se trataba de convertir el lenguaje en una

forma numérica para poder ser procesado por el ordenador. No encontraron difícil esta tarea de conversión del lenguaje y creyeron que los hechos podían ser almacenados en una máquina con suficiente capacidad. El lenguaje tal y como lo concibieron era un sistema finito que obedece a leyes claramente definidas y organizadas lógicamente, y por tanto, susceptibles de análisis cuantitativo. Las pruebas realizadas en estos comienzos fueron un análisis de palabra por palabra, encontrándose inadecuado, ya que, a menudo, para que la frase traducida tenga sentido hay que transformar la estructura, es decir, la naturaleza y, a veces el número de palabras que componen la frase inicial.

En mayo de 1951, Bar-Hillel fue nombrado el primer investigador con dedicación exclusiva para la TA del MIT, y en junio del año siguiente organizó el primer simposio sobre la TA, en el que se reunieron alrededor de 20 investigadores. En dicho simposio se expusieron los avances sobre la TA y se trazaron las líneas a seguir en investigaciones futuras. Quedo claro que para conseguir una traducción de calidad se necesita la intervención humana antes o después del proceso del ordenador, y también la necesidad de profundizar en el aspecto sintáctico. Algunos de los asistentes consideraron que la necesidad más inmediata era demostrar la viabilidad de la TA. Este punto quedo resuelto en enero de 1954 cuando Leon Dostert de la universidad de Georgetown y Paul Garvin (que preparó las bases lingüísticas) hicieron una demostración en la Universidad de Georgetown, utilizando un ordenador IBM 701 (‘International Machines Company’) programado con 250 palabras y seis reglas sintácticas básicas, intentándose traducir del Ruso al Ingles (Pinchuck 1977: 239-240). El programa tradujo con éxito 49 frases escritas en ruso que se habían seleccionado con anterioridad y pertenecientes principalmente al campo de la química orgánica.

El experimento, aunque no tenía un gran valor científico, tuvo un gran éxito y fue la prueba de que la TA tenía futuro. A partir de ese momento comienza lo que algunos investigadores consideran la década del optimismo (1954-1966) (Hutchins 2005) financiándose numerosos proyectos principalmente en las dos superpotencias, los Estados Unidos de América y la Unión Soviética, que dominaban la agenda global de

la política económica, operaciones militares y, progresos científicos incluyendo la iniciación de la exploración espacial. En 1957, con el lanzamiento del primer Sputnik soviético al espacio, la traducción automática se convierte en tema de investigación prioritaria en los Estados Unidos, ya que para alcanzar y superar a los equipos soviéticos era necesario conocer sus resultados y la orientación de su tecnología, lo que llevo a una gran demanda de traducciones del ruso al inglés a fin de poder conocer los contenidos de los documentos científicos y técnicos, tales como artículos que aparecían en revistas científicas rusas. Una traducción en borrador era suficiente para conseguir un conocimiento básico de los artículos. Si el artículo carecía de interés se desechaba y en caso contrario se enviaba a traductores humanos para que completaran la traducción. El Gobierno Estadounidense adjudicó enormes cantidades de dinero a más de cuarenta organismos para que investigaran sobre la traducción automática.

En la Unión Soviética, diferentes instituciones como el Instituto Matemático Steklov de Moscú, el Instituto de Mecánica y Procesamiento de Datos de la Academia de Ciencias y el Instituto de Enseñanza de lenguas extranjeras de Moscú, realizaban investigaciones paralelas.

La TA en aquel momento requería la utilización de máquinas inmensas y lentas que exigían grandes inversiones de dinero. La capacidad de los ordenadores y las teorías del lenguaje no eran bien conocidas, aunque, algunos pretendían que en el curso de unos años se construyesen sistemas capaces de traducir cualquier texto, así que, unos dedicaron sus esfuerzos a la creación de sistemas que resultasen operativos en poco tiempo y otros al estudio de la lingüística. Los primeros sistemas creados, conocidos como "sistemas de primera generación" consistían principalmente en enormes diccionarios bilingües donde las entradas en la lengua fuente daban lugar a uno o más equivalentes en la lengua meta, y algunas reglas para producir el orden correcto de las palabras, si bien con resultados muy limitados.

Pronto se llegó a la conclusión de que el método directo de palabra por palabra no era suficiente y que una traducción de cierta calidad sólo sería posible con un análisis sintáctico y semántico completo tanto de la lengua fuente como de la lengua meta.

Como ya se ha visto, en la década de los cincuenta el optimismo era grande ya que los avances en informática y en la lingüística formal, especialmente en el área de la sintaxis, vaticinaban una mayor calidad en las traducciones. Se predecían grandes avances y sistemas totalmente automáticos que estarían operativos en pocos años, sin embargo, el optimismo inicial iba dando paso a un clima de desilusión entre los investigadores debido a las aparentemente barreras semánticas que se planteaban y la certeza de que los sistemas nunca llegarían a tener el conocimiento del mundo que tiene el ser humano. En 1960 se llegó a las siguientes conclusiones:

a) la necesidad de construir sistemas en los que toda mejora y toda evolución pueda emprenderse fácilmente, es decir la necesidad de separar las gramáticas y los diccionarios por una parte del soporte lógico portador de los algoritmos de análisis y de la traducción, por otra. Los algoritmos ayudaran a analizar las relaciones entre las palabras, y por tanto a comprender el papel de cada palabra, y después transponer estas relaciones en el sistema de la lengua meta. De estos algoritmos dependerá la precisión de la traducción.
b) que no se debe intentar traducir cualquier texto, sino restringir los objetivos a un área determinada.

Ese mismo año Bar-Hillel redactó un informe en el que señalaba que la consecución de un software capaz de realizar traducciones de gran calidad totalmente automáticas estaba lejos de conseguirse y hubo que aceptar que una Fully Automatic High Quality Translation (FAHQT) no era viable a corto plazo debido a la ambigüedad semántica de las palabras y al doble significado de muchas de ellas. Bar Hillel recomendó que los objetivos de la TA deberían ser menos ambiciosos y centrarse en la construcción de un sistema rentable para la interacción entre el hombre y la máquina.

Sin embargo, la investigación sobre la TA continuaba su curso impulsada por intereses surgidos durante la guerra fría. Prueba de ello es que en 1962 existían 48 grupos de trabajo en el mundo que realizaban investigaciones acerca de la TA con el propósito de que los sistemas pasaran textos de una lengua a otra.

En 1964, cuando la decepción era total, el ministerio de defensa de los EE.UU., llegó a destinar 2,5 millones de dólares para la investigación sobre la TA y la Academia Nacional de Ciencias (National Science Foundation) financió a ocho universidades para realizar estudios sobre esta materia. Durante esta época se pusieron en marcha varios sistemas operativos. La fuerza aérea estadounidense utilizaba un sistema operativo, el Mark II, desarrollado por IBM y la Universidad de Washington, mientras que la Comisión de la Energía Atómica utilizaba un sistema desarrollado por la Universidad de Georgetown. Si bien la calidad obtenida con estos sistemas no era buena, sí satisfacía las necesidades de conseguir información rápidamente.

Cinco años después de que el Gobierno Estadounidense promoviera la investigación sobre la traducción automática, no había aparecido ningún sistema, a pesar del enorme gasto que se había hecho. En 1964 el Gobierno Norteamericano solicitó a la National Science Foundation la creación de un comité de evaluación sobre estas investigaciones. Se creó un comité asesor, el ALPAC, compuesto por siete científicos y liderado por John R Pierce para comprobar los progresos realizados en lingüística computacional, en la TA en particular y las expectativas de la TA para el futuro.

En 1966 aparece el famoso informe ALPAC que afirmaba que la traducción automática era más cara, menos precisa y más lenta que un traductor humano por lo que, no tenía ningún futuro inmediato, ni técnico ni económico. Consideraba que no se debía seguir subvencionado la investigación sobre la TA y recomendando el desarrollo de herramientas informáticas para ayudar al traductor humano. Pese a que en algunos sectores el informe ALPAC fue tachado de parcial y falto de futuro, su influencia fue enorme suprimiéndose gran parte de las subvenciones y cesando las investigaciones. Durante diez años la investigación de la TA en EE.UU. estuvo prácticamente extinguida y sólo perduraron investigaciones marginales para la creación de un sistema operacional de traducción del ruso para la NASA.

Como ya hemos dicho, la Unión Soviética realizaba investigaciones paralelas a las estadounidenses. En el resto de Europa, el interés por la TA había aumentado cuando, en el año 1959, se crea la Comunidad Europea para la Energía Atómica (EURATOM). En

esos momentos, Europa era una comunidad multilingüe que necesitaba un método de comunicación rápido. Francia, Italia, Reino Unido, Republica Federal Alemana Suiza, Hungría, Polonia, Bulgaria y Checoslovaquia comenzaron sus estudios sobre la TA.

Otros países como Japón, Canadá e Israel, se sumaron a la investigaciones TA.

1.2. Después del informe ALPAC

La publicación del informe produjo un profundo impacto en la investigación automática en los Estados Unidos y en menor medida en la Unión Soviética. En EE.UU. la investigación sobre la TA prácticamente se abandonó por más de una década. En 1970, el gobierno norteamericano tan sólo financiaba cinco proyectos en las universidades de Berkeley, Wayne y Texas, y en las empresas Talsec (que desarrollaba el sistema de TA llamado Systran) y Logos Development Corporation. De estos cinco proyectos, sólo sobrevivieron los tres últimos. En 1970, el sistema Systran se instaló para la Fuerza Aérea Estadounidense y posteriormente en 1976 fue adquirida por la Comunidad Económica Europea.

La investigación sobre la TA, en esta década, se desarrolló principalmente fuera de Estados Unidos, concretamente en Canadá, Europa occidental, y Japón, respondiendo a distintas necesidades.

La política bicultural desarrollada en Canadá demandaba un elevado número de traducciones inglés-francés que el mercado no podía satisfacer. Por su parte la CEE necesitaba traducir gran cantidad de documentación legal, administrativa, científica y técnica a las lenguas de todos sus miembros.

En 1977 se instaló en Canadá el sistema Meteo para traducir partes meteorológicos del inglés al francés; en ese mismo año la CEE se empezó a trabajar con el programa Systran para traducir documentación del inglés al francés.

Por los ochenta, tanto la diversidad como el número de sistemas de TA, a la vez que el número de países implicados en la investigación aumentaba. Un gran número de sistemas que se basaban en la tecnología computacional estaban en uso. Como resultado de una mayor oferta de

microordenadores y de software de procesadores de texto, se crea la posibilidad de un mercado de sistemas de TA más económicos, y, así Weider Microsoft ofrece el primer sistema de TA que funciona con un ordenador personal. Muchas compañías europeas y estadounidense se aprovecharon de esta posibilidad y así, estos sistemas fueron explotados, por ejemplo, por ALPS y Globalink, entre otras, y japonesas como NEC Y SANYO. Aunque, también, sistemas procedentes de China y Corea consiguieron entrar en el mercado de la TA.

Pero, Japón es, sin duda, en donde se produce una mayor actividad. Con la quinta generación de ordenadores, se supuso que superaría a sus competidores en software y hardware para ordenadores, lo que llevó a muchas grandes empresas como Toshiba, NTT, Brother, Matsushita, Mitsubishi, Sharp, Sayo, Hitachi, NEC, Panasonic, Kodensha, Nova y Oki, a involucrarse en la creación de software para traducir al o del inglés.

En 1988, la Asociación para el Desarrollo de la Industria Electrónica Japonesa (JEIDA), realizó un estudio para determinar el estado de la tecnología de la traducción automática. Como resultado, se publicó el informe JEIDA, que contenía un informe exhaustivo de todos los proyectos sobre la TA en curso o que habían tenido lugar con algún resultado positivo desde 1966. Según JEIDA, se había producido un cambio radical de la situación desde los días de ALPAC, que aconsejaba continuar haciendo estudios especialmente en el campo de la lingüística computacional, en el procesamiento de lenguajes naturales, y en la mejora y desarrollo de datos lingüísticos, tales como gramáticas y diccionarios.

Al final de la década de los ochenta se produce un gran aumento en el número de nuevos métodos de TA. IBM desarrolló un sistema basado en métodos estadísticos. Otros grupos usaron métodos basados en ejemplos de traducción, técnica que ahora se denomina traducción automática basada en ejemplos/example-based machine translation. Ambos métodos se caracterizan por la carencia de reglas sintácticas y semánticas y su dependencia de grandes corpus textuales.

Los comienzos de la década de 1990 se consideran vitales en la evolución de la traducción automática, con un cambio radical en la estrategia de traducción basado en reglas gramaticales. La lengua deja

de percibirse como una entidad regida por normas fijas, y comienza a verse como un conjunto dinámico que cambia en función del uso y de los usuarios, que evoluciona a través del tiempo y se adapta a las realidades sociales y culturales.

La disponibilidad de ordenadores más potentes y más asequibles económicamente produce un gran crecimiento en la TA que paso de ser patrimonio exclusivo de las grandes empresas multinacionales que utilizaban grandes ordenadores (*mainframe*), a ser utilizado en ordenadores personales y estaciones de trabajo para traductores profesionales. La TA, como instrumento de masa, comienza su andadura con la aparición de numerosos programas dirigidos a satisfacer las necesidades de una cada vez más extensa gama de probables usuarios. Las empresas que desarrollaban y vendían sistemas como Systran, Metal, Logos o Futjitsu para los *mainframes* se ven obligadas a reducir sus sistemas, a la vez que mantienen las características de sus productos, para realizar versiones para PCs. Systran–Pro (para uso particular) y Systran-Classic (para uso de traductores), por ejemplo, son versiones basados en Windows del exitoso sistema desarrollado desde la década de 1960 para clientes en todo el mundo. La gran ventaja de ambas versiones es que cuentan con el gran diccionario que ofrece Systran, trabajan en una gran variedad de pares de lenguas en ambas direcciones, y cuentan con un precio asequible. La editorial Langenscheidt adquirió los derechos de venta de la versión de Metal, en colaboración con Gesellschaft für Multilinguale Systeme (GMS). El sistema se llama Langenscheidt T1 y ofrece varias versiones desde el alemán el inglés. También de Alemania es el Personal Translator, productor de IBM y Rheinbaben & Busch, basado en el Logic-Programming Based Machine Translation (LMT) sistema basado en transferencias. Tanto Langenscheidt T1 como Personal Translator están dirigidos a traductores no profesionales, compitiendo, por tanto con otros sistemas del tipo de Globalink.

Con la llegada de Internet y su popularización se abra un nuevo horizonte, debido a su condición globalizadora y plurilingüe, y produce un rápido crecimiento en la traducción automática de aplicaciones directas de Internet (correo electrónico, páginas Web, etc.) donde la necesidad de una respuesta rápida en tiempo real es lo importante,

aunque sea en perjuicio de la calidad. En estos desarrollos, el software de la TA se convierte en un producto de masas, tan familiar como los procesadores de texto, apareciendo productos de software específicamente para la traducción automática de páginas Web. En 1996, algunas compañías empiezan a ofrecer servicios de traducción, a menudo gratuitos, a través de portales Web de traducción automática. Uno de los primeros y bien conocido, fue BabelFish del portal Altavista que ofrecía versiones del SYSTRAN para traducir francés, alemán y español del o al inglés. Igualmente significativo es el uso de la TA para el correo electrónico: el pionero fue CompuServe que, en 1995, introdujo un servicio de prueba basado en el sistema Transcend. Actualmente, Google Language Tools o AltaVista's Babel Fish, ambos usando tecnología Systram, ofrecen servicios de correo electrónico rápido, páginas web, etc., en el idioma deseado, así como la disponibilidad de diccionarios multilingües, enciclopedias, y libre acceso a bases de datos terminológicas.

En los últimos años las investigaciones promovidas por la Unión Europea se centran no tanto en la TA o en el procesamiento de las lenguas naturales como en la creación de herramientas multilingües pensadas para aplicaciones directas, generalmente dentro de un campo restringido. Un ejemplo es el proyecto AVENTINUS que desarrolla un sistema para la policía para el control de drogas y cumplimiento de la ley a fin de que toda la información sobre drogas, delincuentes y criminales esté disponible en una base de datos accesible en cualquiera de las lenguas de la CE.

1.3. La investigación sobre la TA en España

España no ha permanecido ajena al desarrollo de la investigación sobre la TA y ha seguido fielmente el ritmo marcado por la actividad internacional.

Con un repentino interés sobre la TA se inicia la investigación en España en 1985. El punto de partida es la formación de un grupo de investigación y desarrollo por parte de IBM en Madrid. IBM utilizó el Centro de Investigación en Inteligencia Artificial de la Universidad Autónoma de Madrid como sede de un equipo especializado en lenguaje

natural liderado por Luís Sopeña. Este equipo tomó parte primero en el diseño del prototipo MENTOR junto con otro centro en Israel, y más tarde en la adaptación al español de LMT, sistema diseñado en el T.J. Watson Research Center de Estados Unidos.

Poco después, SIEMENS se asentó en Barcelona para el desarrollo del módulo español de su prestigioso sistema Metal. Monserrat Meya, encargada del proyecto, contactó con el filólogo e ingeniero Juan Alberto Alonso y juntos formaron el núcleo de un equipo en el que participaría una lista interminable de colaboradores. Después de 1992 el grupo se constituyó en empresa independiente, INCYTA. Tras un convenio con la Generalidad de Cataluña y la Universidad Autónoma de Barcelona, se desarrolló el módulo catalán.

A finales de 1986 se crearon en Madrid y Barcelona dos nuevos grupos de trabajo entre quienes se repartió el desarrollo de los módulos del sistema EUROTRA, financiado por la CEE. En Madrid el grupo liderado por Francisco Marcos Marín se ocupaba de los aspectos morfológicos y léxicos, mientras que el grupo de Barcelona, liderado por Ramón Cerdá, se ocupaba de las cuestiones de sintaxis y semántica.

Un quinto grupo se formó, en 1987, en los laboratorios de investigación y desarrollo de la empresa Fujitsu en Barcelona. El objetivo era el desarrollo de los módulos de traducción al español del sistema Japonés ATLAS. El grupo estaba liderado por el ingeniero Jorge Vivaldi y los filólogos José Soler y Joseba Abaitua. Esta linea de investigación fue interrumpida en 1992.

Otro grupo dedicado a la TA fue el formado por Isabel Herrero y Elisabeth Nebot de la Universidad de Barcelona que crearon un prototipo de traducción árabe-español en colaboración con Túnez.

De 1993 a 1999 Joseba K. Abaitua Odriozola de la Universidad de Deusto, Arantza Casillas Rubio de la Universidad de Alcalá de Henares y Raquel Martínez Unanue de la Universidad Complutense de Madrid, trabajaron en el proyecto LEGEBIDUNA para el desarrollo de herramientas que aprovecharan los textos de un corpus bilingüe como fuente de datos para la creación de entornos de edición y traducción de documentos administrativos del castellano al euskera con etiquetado SGML/TEI-P3.

Desde 1998, el Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante desarrolla sistemas de TA entre lenguas románicas; estos sistemas son accesibles libremente por Internet: interNOSTRUM, entre el español y el catalán consigue traducir, con buena calidad, del orden de mil palabras por segundo sobre un ordenador estándar; Traductor Universia, entre el español y el portugués, y Apertium, un sistema de TA de código abierto desarrollado en colaboración con un consorcio de empresas y universidades españolas que traduce entre el español y el catalán, el gallego y el portugués.

2. Logros y limitaciones de la traducción automática

Los logros y las limitaciones de la TA van parejos a lo que esperemos de ella.

Si tenemos en consideración a las personas que van a utilizar las traducciones realizadas mediante ordenadores tenemos dos grupos claramente diferenciados. Por un lado, está el punto de vista académico, la rama más purista y la más exigente en lo que a léxico y sintaxis se refiere. Son los traductores y lingüistas que exigen, ante todo que sus traducciones sean textos que no contengan errores de concordancia, expresiones mal traducidas, cambios de sentido, malas interpretaciones o cambios en el orden de las palabras. Es decir, exigen que sus textos sean legibles. Por lo que para este grupo, la TA sólo es adecuada después de una minuciosa post-edición.

Por otro lado, nos encontramos con el punto de vista práctico, personas ajenas al mundo de la lingüística, la gramática y la lengua en general, personas pertenecientes a diversas ramas científicas o técnicas, o empresarios que buscan información para uso interno de sus compañías y que se limitan a leer el texto para obtener las ideas más importantes, sin que les supongan un problema la mala sintaxis o los errores gramaticales. Para estos, es suficiente una traducción borrador en la que se deje entrever el significado y así comprobar si el texto es de interés o no lo es, o para extraer la información precisa del mismo. Incluso cuando el lector conoce la lengua del texto puede resultarle más fácil leerlo en su propia lengua aunque sea una traducción

rudimentaria. Grandes organizaciones como la Comunidad Europea utiliza la TA con este propósito.

Si tenemos en cuenta el tipo de textos, se consideran idóneos para la traducción automática aquellos en los que el lenguaje se mantiene dentro de unos parámetros previsibles como pueden ser los partes meteorológicos, textos jurídicos estereotipados (contratos, normativas internacionales, etc.), disposiciones legales y administrativos (boletines oficiales, resoluciones etc.), boletines informativos (bolsa, anuncios por palabras, teletexto, ofertas de empleo, etc.), manuales técnicos, resúmenes de publicaciones técnicas, textos científicos no creativos, correo electrónico, consultas a bases de datos en otras lenguas.

Sin embargo, no se consideran apropiados los textos creativos con un lenguaje elaborado y estilizado como poesía, teatro, ensayos filosóficos, críticas, reseñas, etc., y textos expresivos tales como el lenguaje coloquial, los juegos de palabras, o el lenguaje periodístico.

Hay dos ventajas evidentes de la TA: por un lado, aumentan la productividad ya que un traductor humano sólo puede procesar 4-6 páginas ó 2.000 palabras por día para lograr una buena traducción, mientras que utilizando la traducción automática con *post edición* o utilizando herramientas de asistencia al traductor que no se encuentran integradas en un único sistema o entorno, por ejemplo, bancos de datos terminológicos como el EURODICAUTOM, esta cifra se supera grandemente. Por otro lado, se crea una homogeneidad de términos y de estilo.

El principal problema de la TA es la ambigüedad, debido a que los ordenadores no cuentan con un sistema de comprensión como el que tiene el ser humano. El sistema sólo puede eliminar ambigüedades de acuerdo con su configuración, es decir, de acuerdo con lo que contenga su diccionario y los medios de decodificación de una lengua y remodificación en otra lengua. Las ambigüedades son de varios tipos y se pueden solucionar de distintas maneras:

1. La *ambigüedad léxica* se presenta cuando una palabra puede tener más de una interpretación. Puede ser de tres tipos: ambigüedad léxica categorial, ambigüedad léxica: homografía y polisemia y ambigüedad léxica de transferencia o de traducción:

a) La *ambigüedad léxica categorial* es cuando se da la posibilidad de asignar a una palabra más de una categoría gramatical o sintáctica dependiendo (por ejemplo, sustantivo, verbo o adjetivo) del contexto. Un ejemplo son las palabras españolas *vino* que puede ser un sustantivo *o* verbo, o *como* que puede ser un verbo o adverbio. A menudo puede resolverse atendiendo a la flexión morfológica y mediante el análisis sintáctico. Otro ejemplo lo encontramos en la frase inglesa *Gas pump prices rose last time oil stocks fell* en que cada palabra de la oración tiene al menos una ambigüedad categorial (sustantivo o verbo e incluso la palabra *last* puede ser sustantivo, verbo, adjetivo y adverbio). Sólo hay un modo de analizar correctamente esta oración y requiere de un análisis sintáctico en profundidad.
b) La *ambigüedad léxica: homografía y polisemia* es cuando una palabra tiene dos o más posibles significados diferentes. Dos o más palabras son homógrafas cuando tienen la misma forma pero significados diferentes y son polisémicas si muestran una variedad de significados relacionados de algún modo entre sí. Ambas pueden recibir el mismo tratamiento en un análisis de TA. Para eliminar este tipo de ambigüedad hay en muchos casos que esperar a que el resto de la oración proporcione más información lingüística o contextual.
c) La *ambigüedad léxica de transferencia o de traducción* se produce cuando una palabra de la lengua de origen puede traducirse a diversas palabras o expresiones en la mengua meta. La ambigüedad no se produce con respecto a la lengua de origen sino con respecto a la traducción.

2. La *ambigüedad estructural* es cuando una misma oración puede interpretarse mediante distintas estructuras sintácticas. Surge cuando la estructura profunda de una oración puede analizarse de más de un modo según esté definida la gramática empleada por el sistema. Se deben a combinaciones accidentales de palabras que tienen ambigüedades categoriales, usos gramaticales alternativos de los constituyentes sintácticos y a distintas combinaciones posibles de los constituyentes. La ambigüedad estructural difiere de una lengua a otra y dentro de una lengua de una gramática a otra. La solución

es escoger una de las posibles interpretaciones de una oración por lo que la traducción puede variar según la interpretación escogida en la lengua de origen. La mayoría de las ambigüedades estructurales podrían resolverse con información contextual, pero son escasos los sistemas que lo utilizan:

a) por no haber reglas para definir dónde buscar la porción de conocimiento necesario,
b) por el tiempo necesario para almacenar la información y cual sería la vigencia del conocimiento extraído y
c) por el coste computacional asociado.

Claro que se puede recurrir a otras estrategias como seleccionar la estructura más probable, si el sistema de TA es interactivo puede pedir al usuario que seleccione él la interpretación, o si las lenguas tienen una estructura y vocabularios parecida se puede recurrir al pase gratuito o *free ride*, es decir, no resolver la ambigüedad y mantenerla como tal en la lengua meta.

3. La *anáfora* es la referencia indirecta a una entidad mencionada de forma explícita en otro lugar del texto. Los recursos lingüísticos que se utilizan para realizar una referencia indirecta son los pronombres, demostrativos o expresiones como *el último*, *el anterior*, etc. En muchos casos es importante identificar el antecedente (el objeto al que se refiere la referencia indirecta) de la anáfora para traducir correctamente. Se requiere el mismo tipo de conocimiento, lingüístico, contextual y del mundo real, que para despejar otros tipos de ambigüedad.
4. Finalmente, la *ambigüedad en el alcance de los cuantificadores* se produce cuando cuantificadores como *alguno*, *ninguno*, *todos* son imprecisos y para resolverlos es necesario conocimiento contextual y del mundo real.

Aparte de la ambigüedad, hay otros aspectos que generan problemas a la hora de realizar una traducción con un sistema informático, como por ejemplo, los juegos de palabras o los modismos. Por ello, se ha tendido a limitar el campo del texto ya que cuanto más simple sea la gramática y la sintaxis de texto de origen y más restringido sea el texto, mejores serán los resultados.

También es difícil que un programa de traducción pueda determinar la intención de un texto, ya que carece de la capacidad de juzgar.

Por todas estas razones, es muy frecuente que se tenga que acudir a la intervención humana si se quiere que la calidad de la traducción sea mayor.

3. Niveles de la traducción automática

En un principio, la TA surgió con la idea utópica de obtener sistemas de traducción totalmente automáticos y con resultados de alta calidad, comparables al mejor traductor humano. Los problemas que fueron surgiendo, dirigieron las investigaciones y los desarrollos hacia sistemas capaces de obtener traducciones de suficiente calidad para las necesidades del usuario o bien sistemas de ayuda a los traductores profesionales La traducción automática engloba numerosos sistemas que sólo comparten la utilización del ordenador como instrumento de traducción.

Dependiendo de si hay intervención humana o no en el proceso de la traducción tenemos: traducción automática sin intervención humana y traducción realizada con el ordenador con intervención humana.

3.1. Traducción automática sin intervención humana

Siempre se ha perseguido la traducción totalmente automático, sin intervención humana y de gran calidad (FAHQT), pero hasta ahora no se ha logrado.

Aunque se ha visto como objetivo utópico el conseguir programas de FAHQT, y los investigadores no desechan este ideal, lo cierto es que sí es posible obtener programas de traducción de alta calidad acotando la entrada de los mismos, bien mediante preedición para preparar el texto origen para adaptarlo a las exigencias del programa, eliminando así las ambigüedades, o bien sea con la utilización de sublenguajes.

3.2. Traducción automática con intervención humana

Los CAT incluyen tanto la traducción automática asistida por humanos (Human Aided Machine Translation [HAMT]) como la traducción humana asistida por ordenador (Machine Aided Human Translation [MAHT]).

Mediante la interacción con el sistema en el proceso de traducción, o mediante la postedición de los resultados, se puede obtener una traducción de alta calidad.

3.2.1. Traducción automática asistida por humano

En la traducción por ordenador asistida por el hombre (HAMT), el ordenador lleva la iniciativa, permitiendo al usuario desprenderse de las tareas más pesadas como puede ser la búsqueda en los diccionarios. El usuario puede intervenir antes, durante o después del trabajo. Es un sistema interactivo. En el estado de pre-edición modifica las frases a traducir, mientras que durante el proceso, si el ordenador encuentra alguna parte complicada entabla un dialogo con el usuario a fin de resolver ambigüedades, seleccionar el término más adecuado, identificar referentes, etc., y en el fase de post edición, corrige los resultados de la traducción. Los sistemas ALPS (Comercializados en Francia por Control Data) y Weidner, son ejemplos de este tipo en los que se ha insistido en el papel de la intervención humana.

3.2.2. Traducción humana asistida por ordenador

En la traducción humana asistida por ordenador (MAHT), el responsable es el ser humano que emplea la máquina para el acceso y consulta on-line de diccionarios (generales y específicos), acceso a bancos de terminología, repertorios de frases estereotipadas, clichés, ejemplos de uso de palabras, etc.

El sistema MAHT es usado por la Oficina Federal de Idiomas de Alemania, la Comunidad Europea, el Gobierno Canadiense, etc. Según un estudio hecho por *The Federal Armed Forces Translation Agency* en Manheim se comprobó que el trabajo de un traductor con ayudas convencionales requería entre el 50 y el 80% más de tiempo que un

traductor trabajando en un glosario computerizado. Además estos últimos cometen de uno a tres errores menos.

4. Tecnología de los sistemas

Una clasificación desde el punto de vista metodológico, es decir como abordan los programas el problema de la TA, diferencia entra dos grandes grupos: los que se basan en reglas lingüísticas por una parte, y los que utilizan corpus textuales por otra.

4.1. Sistemas basados en reglas

La traducción automática mediante reglas consiste en realizar transformaciones a partir del original, reemplazando las palabras por su equivalente más apropiado.

En general, en una primera fase se analizará un texto, normalmente creando una representación simbólica interna. Dependiendo de la abstracción de ésta representación se encuentran tres grados diferentes: *traducción directa*, *traducción por transferencia* y *traducción interlingual*.

4.1.1. Traducción directa

La traducción se realiza reemplazando palabras de la lengua de origen por palabras en la lengua meta. Para realizar la traducción utilizan como principal recurso las listas de palabras o grupos de palabras (diccionarios) en la lengua de origen y las correspondientes palabras en la lengua meta.

Se concibe el texto como un conjunto de oraciones independientes y las oraciones a su vez como un conjunto de palabras invertebradas. Apenas existe análisis sintáctico y tienen una nula capacidad semántica, además, al no efectuar un verdadero análisis sintáctico no son capaces de detectar ciertas informaciones gramaticales importantes, ni de identificar información semántica, ni detectar palabras con múltiples sentidos.

Cada par de idiomas se trata de manera independiente y no es posible generalizar resultados. Pero, a pesar de las limitaciones, algunos de estos sistemas han demostrado ser rentables y útiles, especialmente cuando nos hallamos ante un texto con un vocabulario y un estilo que están bien definidos y limitados. La calidad de la traducción está en relación directa con la calidad de la información que contiene el diccionario y en general, producen traducciones aceptables y legibles, aunque con elevada dosis de post-edición.

La han usado: Systran, en su primera etapa, utilizado por CEE, el grupo GAT de la Universidad de Georgetown; Spanam utilizado por la *Pan-American Health Organization*; GTS; y PC-Translator entre otros.

4.1.2. Sistemas de transferencia

En la traducción por transferencia, el análisis del original juega un papel más importante, y da paso a una representación interna que es la que se utiliza como enlace para traducir entre lenguas distintas. El proceso de traducción se divide en tres fases: *análisis*, *transferencia* y *generación*. Primero, se lleva a cabo el *análisis* de la lengua de origen para conseguir una representación intermedia con información sintáctica y/o semántica. Esta fase se realiza con un programa que interpreta la secuencia de caracteres de la oración de entrada y construye una representación de la estructura sintáctica de la misma.

A continuación está la fase de *transferencia* de la lengua de origen a la lengua meta. Esta fase se puede producir en varios niveles: léxico, semántico y sintáctico. La transferencia léxica o traducción de los términos que forman la oración de entrada se realiza a partir de la información contenida en el diccionario; la transferencia semántica es la transformación de representaciones profundas: patrones semánticos; transferencia sintáctica o aplicación de una serie de reglas de transformación a la estructura surgida en la fase de análisis, para conseguir una estructura equivalente en la lengua meta.

Finalmente, la fase llamada *generación*, ya que se reconstruye el texto de la oración en la lengua meta, a partir de la estructura obtenida en la transferencia.

Este sistema permite que el procesamiento de la lengua de origen y la lengua meta sean independientes.

La dificultad de este método la encontramos en el desarrollo de sistemas multilingües, ya que, hay que diseñar fases de transferencia para cada par de lenguas y sentido de la traducción. Si queremos resaltar la ventaja de este sistema diremos que el diseño de la fase de transferencia no es muy complejo ya que la representación intermedia es una abstracción dependiente de la lengua.

Ejemplos de sistemas de transferencia son: Systran en su segunda etapa; Metal de la Universidad de Texas; GETA de la Universidad de Grenoble; Taum Meteo de la Universidad de Montreal; EUROTRA de la CEE, Logos, etc.

4.1.3. Sistemas interlingua

Los primeros sistemas interlingua fueron propiciados por los teóricos de la inteligencia artificial (AI) con sus modelos de representación del conocimiento. La tesis que mantiene este método es que para traducir un texto hay que comprenderlo previamente.

La traducción se realiza basándose en una representación conceptual intermedia conocida por interlingua o lenguaje pívot.

En este sistema hay dos componentes monolingües: el análisis de la lengua de origen a interlingua, y la generación de interlingua a lengua meta. La información semántica se recoge en una base de conocimientos accesible durante el proceso de traducción.

La dificultad de este sistema radica en la definición de una representación interlingua universal que pueda ser una representación intermedia entre cualesquiera lenguas. Las ventajas son: por un lado, facilita el desarrollo de sistemas multilingües ya que el módulo de análisis es independiente del de generación, y, por otro, incorpora los niveles de análisis lingüísticos de las distintas lenguas.

Entre los proyectos interlingua cabe destacar la implementación realizada por el equipo de Sergei Nirenburg de la Universidad de Carnegei Mellon; o los realizados en la Universidad de Southern California. Pero, sin lugar a dudas, el proyecto más ambicioso es del consorcio asiático *Centro para la Cooperación Internacional e Informatización* (CICC), con grupos de trabajo en Japón, China,

Tailandia, Malasia e Indonesia. Este proyecto está promovido por el Gobierno Japonés y realizado con el proyecto de diccionario electrónico ERD.

4.2. Sistemas basados en corpus textuales

A partir de la década de los 90, aplicando técnicas similares a las utilizadas en Lingüística de Corpus y Reconocimiento del Habla, aparecen una serie de investigaciones basadas en métodos de aproximación estadística sobre muestras de textos previamente traducidos. Son sistemas llamados de traducción asistida.

El que las investigaciones con métodos probabilísticos consigan buenos resultados se debe a la rápida evolución y abaratamiento del hardware, así como, a la aparición de Internet y con ella la posibilidad de disponer de un mayor número de textos en formato digital. La traducción automática a partir de corpus lingüísticos se basa en el análisis de muestras reales con sus respectivas traducciones. Entre los sistemas que incluyen corpus encontramos: *métodos estadísticos* y *traducción mediante ejemplos*.

4.2.1. Métodos estadísticos

Su objetivo es generar traducciones a partir de métodos estadísticos basados en corpus de textos bilingües de millones de palabras, La táctica consiste en calcular los parámetros para el modelo de traducción utilizando técnicas de alineamiento (es decir, haciendo explicitas las correspondencias entre segmentos del corpus bilingüe) de frases en los dos idiomas y calculando la probabilidad de que una palabra corresponda a ninguna, una o más de una palabra en la lengua meta.

Los métodos estadísticos no significan un cambio de estrategia en la traducción, sino en el desarrollo de los componentes, fundamentalmente en los diccionarios de correspondencias de la lengua original y la lengua meta.

Son considerados lentos, sin embargo, los experimentos realizados dan un alto porcentaje de aciertos. Recientemente han mejorado sus capacidades traductoras al añadir 200 billones de palabras de las Naciones Unidas.

Cabe destacar el proyecto *Candice* desarrollado por IBM.

4.2.2. Traducción por ejemplos

Este enfoque coincide con el de la técnica basada en estadísticas, ya que se fundamenta en datos tomados de un corpus bilingüe como principal fuente de conocimiento.

Este método funciona extrayendo oraciones de un corpus bilingüe previamente alineado, utilizando la traducción de los mismos como modelo de nuevas traducciones. Podríamos decir que es la reutilización de traducciones humanas que han sido validadas después de haber sido analizadas.

Suele aplicarse como traducción asistida, en donde el ordenador interacciona con el usuario proponiendo ejemplos de traducciones.

Entre los investigadores que han propuesto este método encontramos a Nagao y Sato de la Universidad de Kyoto, Sadle de ATR, Saito y Tomita de CMU, Somers, Tsuji y Jones de UMIST.

4.3. Sistemas actuales

La tendencia actual en la arquitectura de sistemas es la integración de componentes con varias metodologías (lingüísticas, estadísticas, u otras), a la base de datos de un corpus, dando lugar a diferentes sistemas híbridos. La elección del enfoque se realizara según sea la aplicación.

También están los gestores de memorias de traducción que emplean las técnicas de traducción basada en analogías para utilizar traducciones previamente validadas por un traductor humano y almacenadas para la posterior traducción de textos similares, siendo recuperados de manera automática ayudando a la labor de traducción. Permite traducir automáticamente frases exactamente iguales a las almacenadas en la memoria y, además, ofrecen frases similares como sugerencia para hacer la traducción. Las frases sugeridas pueden confirmarse sin más, siendo sustituidas automáticamente en la traducción o pueden ser copiadas y modificadas por el traductor. Si el contenido de las memorias es adecuado, y el sistema de indexado y recuperación es correcto, la productividad del traductor se incrementa y resultan de gran ayuda en el proceso de traducción.

Permite crear memorias durante la fase de traducción o sucesivamente, si se tiene una versión de la lengua en cuestión y de la versión correspondiente ya traducida. Se pueden crear memorias diferentes según el contexto, y utilizar varias memorias para analizar automáticamente los textos.

Esta tecnología se ha llevado al mercado con un considerable éxito en paquetes que incluyen diversas herramientas de apoyo al traductor. Entre las más conocidas tenemos TRADOS (Translator's Workbench, DEJA-VU (ATRIL), WORDFAST, TRANSIT (STAR), etc.

5. Los Sistemas más significativos

Actualmente hay muchos programas de software para traducir el lenguaje natural, muchos de ellos *on-line*, como Systran, utilizado tanto por Google como por Alta Vista's Babel Fish. Aunque ningún sistema proporciona la tan buscada FAHQMT, muchos de ellos producen una traducción razonablemente buena.

Para completar este estudio de la TA haremos un repaso a algunos de los sistemas que han existido y que existen. Por supuesto, se trata de una selección, ya que ofrecer una visión completa de todos los sistemas que ha habido y hay es prácticamente imposible. Al tratarse de una selección, quizá, para muchos interesados en la TA no sea todo lo extensa que debería ser, pero sí servirá para que el lector se haga una idea general.

5.1. El sistema Systran

Cuando las investigaciones sobre la traducción automática se encontraban casi en punto muerto en todas partes, a consecuencia del ana1isis desfavorable emitido por el, 'Informe ALPAC', un equipo de la Universidad de Georgetown, Estados Unidos, dirigido por Peter Toma, elaboró el Systran, uno de los primeros sistemas de traducción automática. Su presentación se realizó en 1970.

Se concibió en un marco militar, con vistas a traducir documentos científicos y técnicos del ruso al inglés y viceversa para las fuerzas

aéreas americanas. Systran es mundialmente conocida al ser utilizado en el curso de la misión Apollo-Soyouz en 1975 para traducir documentos de importancia vital del ruso al inglés y viceversa.

El sistema se introdujo en 1976 en Luxemburgo con la compra de la versión que traducía del inglés al francés; en 1977 compró la versión francés-inglés y en 1978 la versión inglés-italiano. Desde entonces, la Comunidad Europea ha financiado su desarrollo en las lenguas de sus países miembros. Desde su introducción los ingleses M. Masterman y V. Lawson no ahorraron esfuerzos para hacer utilizable el sistema y mejorar la calidad de la traducción. En el año 1980 este programa tradujo más de 20 millones de palabras. Poco a poco se ha ido extendiendo por todas las lenguas europeas (alemán, español, francés, italiano, portugués) y han sido necesarios muchos años para poner a punto los diccionarios y gramáticas en cada una de estas lenguas.

Su funcionamiento es el siguiente: antes del tratamiento propiamente dicho, se efectúa un control ortográfico del texto fuente, requiriéndose la asistencia de un operador que corrija y complete las palabras no reconocidas por la máquina. El texto fuente es entonces presentado en una pantalla terminal, y el operador elige la lengua a la que desea traducir, así como tres glosarios, citados por orden de prioridad, a fin de que el programa terminal mejore la traducción de las palabras.

La secuencia de las operaciones que efectúa se divide en tres grandes etapas: a) la fase de ana1isis que comienza con la reso1ucion de las homografías; siguiendo con los nombres compuestos, la identificación de frases, las relaciones sintácticas primarias, las estructuras de coordinación, identificación de sujeto/predicado para terminar con las identificación de las estructuras de tipo preposicional; b) la etapa de transferencia que pasa por la resolución de ambigüedades, es decir, corresponde a grandes rasgos a la traducción de las categorías y de las estructuras analizadas; c) la tercera parte, la generación, es la traducción de las palabras en si misma.

Grandes colosos de la industria norteamericana, como Xerox Corporation lo utilizan para traducir manuales de mantenimiento al francés, español, portugués, italiano y alemán y General Motors y Ford Motors para la traducción de documentos técnicos de inglés a

francés. Según la documentación, las traducciones se procesan a una velocidad de 500.000 palabras por hora.

Actualmente Systran proporciona tecnología a Yahoo!, AltaVista's (Babel Fish), a la mayoría de los lenguajes ofrecidos por Google's language tools.

Existen versiones comerciales que corren bajo Microsoft Windows (incluyendo Windows Mobile), Linux y Solaris.

5.2. Meteo

Este sistema fue desarrollado por John Chandious para el Gobierno Canadiense. La versión Meteo 1 fue puesta en marcha en marzo de 1977 y desde entonces ha estado funcionando las 24 horas del día hasta que el 30 de septiembre de 2001 fue reemplazado por un sistema competidor. El sistema ha sufrido distintas mejoras a lo largo de su historia. La versión Meteo 2 empezó a funcionar en 1983. El software corría con 48Kb de memoria central y 5 Mb de disco duro. Se considera la primera aplicación de TA que ha funcionado con microordenadores. En 1996, se desarrolló la versión Meteo 96 que se usó para traducir los partes meteorológicos durante lo Juegos Olimpcos de Atlanta. La última versión conocida de este sistema es Meteo 5 de 1997. Operaba sobre un IBM PC network con Windows NT. Traducía 10 páginas por segundo, ocupando muy poco espacio, pudiéndose meter en un disquete de 1.44 MB.

Su objetivo estaba bien definido: traducir al francés los boletines Meteorológicos emitidos por aviones, globos sonda, estaciones terrestres, etc., y que estaban redactados en inglés. Seguramente su éxito se debe a los límites operativos; esta clase de boletines se basan a menudo en las mismas palabras y las mismas expresiones, con las que es fácil establecer un diccionario. La estructura de la frase es también sencilla y no presenta ambigüedades. Con un diccionario pequeño de unas 2.000 palabras y expresiones, el sistema puede traducir y si la máquina queda encallada en una frase, ésta es presentada en una pantalla y traducida por un traductor humano. El Meteo es un ejemplo clásico de sistema eficaz y apropiado para una tarea de un dominio limitado, además de resultar económicamente rentable ya que el 80% de los

boletines son traducidos sin intervención humana. Estaba traduciendo unas 80.000 palabras al día o 30 millones de palabras al año.

Al igual que el Systran opera en tres etapas: a) análisis de una edición previa del texto fuente, un análisis morfológico, una consulta de diccionario y un análisis sintáctico; b) transferencia primero léxica y después sintáctica; y c) generación en la que primero es la sintáctica y después la morfológica.

5.3. Titran

Es un sistema que traduce títulos de artículos científicos y referencias bibliográficas. En 1981 apareció el Titran EJ (inglés/japonés) y un año mas tarde el Titran JE (Japonés/inglés) y por último y gracias a los trabajos de Jerome Huber, existe el Titran JF (japonés/francés).

El soporte lógico utilizado en las dos últimas versiones es el mismo, lo que prueba la aptitud de las máquinas de traducir textos en un dominio limitado y con estructuras sintácticas poco complejas. El diccionario está constituido en gran parte por palabras sencillas y vocabulario técnico. El primer trabajo del sistema es descomponer las palabras del texto japonés que se presenta en forma de una serie de caracteres, después de esta segmentación del texto original y de la identificación de sus grupos sintácticos, el programa procede a su transferencia gracias a un conjunto de reglas de gramática escritas en un 'Metalenguaje'.

El Titran EJ se ha utilizado para traducir los títulos de la base de datos bibliográficos INSPEC en los campos de la física y la informática. El 90% de los títulos son traducidos automáticamente por el sistema de forma correcta. El Titran JE y JF sólo se han experimentado para la informática y su tasa de traducción es también del 90%.

5.4. Logos

Logos Corporation fue fundada por Bernard Scott en 1970 y durante treinta años trabajó en el sistema Logos, hasta que en el 2000 la compañía se disolvió. Empezó ofreciendo traducciones del vietnamita al inglés. Actualmente, Deutsche Forschungszentrum für Künstliche

Intelligenz (DFKI) ha desarrollado una versión abierta del original Logos, y se ha convertido en uno de los competidores de las grandes sistemas como Systram o IBM's WebSphere. Hay que destacar su precisión y el hecho de no descartar la necesidad de un traductor humano, además admite diferentes formatos de documentos y mantiene el formato del documento original en la traducción. Incluye los pares alemán al inglés y francés, e inglés a las más importantes lenguas Europeas (francés, italiano, portugués y español).

La nueva versión de Open Logos trabaja con Linux y esta disponible, completamente libre, para traductores, universidades o cualquier persona.

Su velocidad puede ser de hasta 100 páginas en una hora, es decir 320.000 palabras. Entre sus clientes se encuentran IBM Germany, Hewllett Packard, o Nixdorf.

5.5. Spanam y Engspan

Son los sistemas utilizados por la Organización de la Salud Panamericana con sede en Washington Dc. Cuenta con más de 2.000 traductores en plantilla, pero dado el inmenso volumen de traducciones al que se enfrenta, ha optado por dos sistemas de traducción automática: Spanam, traduce del español al inglés y comenzó en 1979; y Engspan introducido en 1984 para traducir del inglés al español.

5.6. Otros sistemas

Otros muchos sistemas han aparecido, entre los más significativos está el Ariane-78 en el que no es la palabra si no la frase considerada como un todo la que se traspasa de la lengua de origen a la lengua meta. Este sistema fue desarrollado por el Grupo de Estudios para la Traducción Automática (GETA) de Grenoble, que dirige el profesor Bernard Vanquois.

Según el propio Bernard Vanquois es un sistema de la segunda generación, ya que separa netamente el soporte lógico del sistema que incluye los algoritmos de la traducción, y los modelos lingüísticos que incluyen los diccionarios y las gramáticas propias de cada idioma,

haciendo su funcionamiento mas fácil de prever y más fácil de mejorar que en un sistema del tipo Systran.

Ariane-78 establece una estructura arborescente, donde las diferentes "unidades léxicas" llevan unas etiquetas que precisan sus características morfológicas, sintácticas y, en cierta medida, semánticas. Esta representación normaliza la organización sintáctica de la lengua antes de efectuar la traducción propiamente dicha. Esta transferencia conduce al mismo tiempo a la estructura de conjunto (arborescencia) y a las unidades relativas al léxico, es decir, a las palabras definidas de forma muy general, sea cual sea su forma (nombre, verbo, adjetivo, etc.) en que aparecen. Una vez realizada esta doble transferencia el sistema genera seguidamente la frase completa en la lengua meta, recorriendo en sentido inverso las mismas etapas que en la primera fase del análisis.

Otro sistema es el Weidner, creado por Bruce Weider en la Universidad de Brigham Young en 1977. Se comercializó en Europa en 1980 por TAO Internacional con el nombre de CAT. Es más bien un sistema de traducción asistida por ordenador que de traducción automática. Existen dos versiones: el Micro CAT que es el primer sistema de traducción automática que funciona con un ordenador personal. En sus inicios el par de lenguas con las que funcionaba eran inglés-francés, y luego se añadieron inglés-español e inglés-alemán. Actualmente trabaja con 7 pares de lenguas (además de los ya mencionados están: español-inglés, inglés-portugués, francés-inglés, e inglés-italiano). En sus comienzos generaba, a una velocidad de 1.800 palabras a la hora, traducciones toscas y ahora son de 4.000 a 8.000 palabras hora.

La otra versión es el Macro CAT que se concibió para que numerosos traductores trabajando simultáneamente en varias lenguas pudieran utilizarlo. Mucho más rápido que el Micro CAT, genera traducciones a una velocidad que puede alcanzar los 8.000 palabras a la hora. Lo mismo que para el Systran, cada pareja de lenguas necesita un soporte lógico distinto. Actualmente funciona en alemán/inglés, inglés/árabe y japonés/inglés además de las mismas parejas de lenguas en las que funciona el Micro CAT.

Entre sus clientes se encuentra Aérospatiale, Bull, Matra, Télésystemes y Thomson en Francia y Perkins Engines en el Reino Unido.

En 1982, se inicia el proyecto Eurotra lanzado por la Comunidad Europea, como sistema moderno de TA, que a largo plazo, se esperaba que desembocara en la construcción de un sistema multilingüe en siete idiomas. El objetivo era proporcionar un análisis de un texto original en una de las lenguas de la Comunidad de modo que sirviera como base para la transferencia y la generación de un texto en cualquier otra lengua de la comunidad. Pero en 1991, tras el "Informe Danzan", se da por concluido dicho proyecto con cierta sensación de fracaso.

En 1970 el Institut textile de France desarrolló el Titus que suministra, por cada documento pedido un resumen simultaneo en alemán, inglés, castellano y francés. Traduce entre francés, inglés, alemán y español. Este sistema tiene no solamente la ventaja de ser multilingüe –el número de lenguas no esta limitado, sino sólo restringido a la familia de lenguas indoeuropeas– sino que es también muy económico de memorizar, ya que una frase de cuarenta caracteres alfabéticos se condensa en tres caracteres-máquina. Requiere una pre-edición considerable. Un editor analiza la frase y de forma interactiva corrige errores y ambigüedades.

En la década de los noventa, encontramos proyectos industriales, como METAL, que comenzó en Texas y fue adquirido e impulsado por la empresa Siemens. Funciona con lenguas alemán-inglés, francés y holandés y alemán-español.

En 1980 aparece el ALPS que, al igual que el sistema Weider, procede de la Universidad de Brigham Young. Ofrece pares del inglés al francés, alemán y español; del francés al inglés, y del alemán al inglés. Permite un alto nivel de prestaciones que lo hace interactivo. Sus clientes son Texas Instruments, Unisys, NCR France, OTAN, Norsk Data y las empresas en la red Alpnet.

PC Translator de Linguistic Products, comenzó con el danés-inglés y posteriormente se añadieron inglés francés e inglés-español. El sistema sólo es útil para traducciones repetitivas.

Globalink. Esta empresa ha emprendido una ambiciosa campaña de adquisición de productos, entre los que destaca Language Assistant.

Conclusiones

Los avances en la tecnología de la información (TI) se han combinado con las necesidades modernas de comunicación para fomentar la traducción automática. La relación entre la tecnología y la traducción se remonta a los inicios de la Guerra Fría, en el decenio de 1950, la competencia entre los Estados Unidos y la Unión Soviética era tan intensiva en todos los niveles, que miles de documentos fueron traducidos del ruso al inglés y viceversa. Aunque la guerra fría ha terminado y, a pesar de la importancia de la globalización, que tiende a romper barreras culturales, económicos y lingüísticas, la traducción sigue siendo una actividad primordial, debido al deseo por parte de las naciones de mantener su independencia y su identidad cultural, especialmente el poder expresarse en su propia lengua. Este fenómeno se puede ver claramente en la Unión Europea, donde la traducción sigue siendo una actividad fundamental.

La existencia de numerosos programas de ordenador capaces de traducir una gran variedad de textos de una lengua a otra es ya un hecho y no una quimera. Pero no existen máquinas de traducir que sean capaces de analizar textos en cualquier idioma y producir una traducción perfecta en otro idioma sin intervención humana. Hoy existen numerosos programas de software, incluso varios de ellos *on-line*, para traducir lenguajes naturales que proporcionan una relativamente buena traducción, pero aún no existe ningún sistema que proporcione la tan buscada *Fully Automatic High Quality Translation*, que sigue siendo es una aspiración para el futuro y sólo en ese futuro sabremos sí se puede realizar. Lo que sí se ha logrado hasta el momento es desarrollar programas que realizan traducciones con una calidad relativamente buena de textos en áreas bien definidas, como se desprende de las palabras de la Asociación Europea para la TA (EAMT):

> Machine translation (MT) is the application of computers to the task of translating texts from one natural language to another of the very earliest pursuits in computer science, MT has proved to be an elusive goal, but today a number of systems are available

which produce output which, if not perfect, is of sufficient quality to be useful in a number of specific domains' (1997).

Un factor determinante en la calidad de la traducción es el grado de especialización de los sistemas de traducción, que mejoran en la medida en que se adaptan al tipo de texto y vocabulario que se vaya a traducir. Un ejemplo claro es el Meteo que consigue altas cotas de calidad pero no se puede utilizar para traducir documentos económicos.

De todas formas, ya no se piensa en la traducción automática como un sustituto del traductor humano sino que se cree que para que una traducción sea perfecta al cien por cien es necesaria la cooperación de ambos. Así, actualmente el software relacionado con la traducción, se diseña principalmente como una ayuda al traductor humano. Es la llamada traducción ayudada por ordenador. Las CAT proporcionan una gran variedad de herramientas lingüísticas para mejorar la productividad de los traductores, especialmente cuando la traducción es altamente repetitiva, como la documentación técnica.

La investigación continúa tanto en ámbitos académicos como en empresas de software dedicadas al desarrollo de entornos de traducción. Los avances en la tecnología de los ordenadores, en lingüística teórica y en inteligencia artificial, así como la constante búsqueda de herramientas válidas para agilizar el trabajo de los profesionales del sector, están marcando las futuras líneas de investigación.

Referencias

Abaitua, Joseba. 1997. *Traducción automática: Presente y Futuro*. http://www.foreignword.com/es/Technology/art/Abaitua/Abaitua_1.htm Consultada 6 de noviembre de 2007.

Abaitua, Joseba. 2002. *Introducción a la traducción automática (en 10 horas)*. http://paginaspersonales.deusto.es/abaitua/konzeptu/ta/mt10h-es Consultada 21 de noviembre de 2007.

ALPAC. http://en.wikipedia.org/wiki/ALPAC_report
Ultima modificación 12 de octubre de 2007. Consultada 6 de noviembre de 2007.

European Association for Machine Translation. 1997. *What is a translation?*
http://www.eamt.or/mt.html
Última modificación 3 de junio de 2004. Consultada 6 de noviembre de 2007.

Hernández, Pilar. 2002. *En torno a la traducción automática*.
http://internet.cervantes.es/internetcentros/cultura/pdf/presentacion2.pdf
Consultada 6 de noviembre de 2007.

Hutchins, W.J. 1986. *Machine Translation: Past, Present, Future*. Hichester: Ellis Horwood.

Hutchins, W.J. 1995a. *Introducción a la traducción automática*. Madrid: Visor.

Hutchins, W.J. 1995b. "Machine Translation: A Brief History". In *Concise History of the Language Sciences: From the Sumerians to the Cognitivists*, E.F.K. Koerner y R.E. Asher (eds.). Oxford: Pergamon Press. 431-445.

Hutchins, W.J. 1999a. "Milestones in Machine Translation nº6: Bar-Hillel and the Nonfeasibility of FAHQT". *International Journal of Language and Documentation* 1.

Hutchins, W.J. 1999b. *The Development and Use of Machine Translation Systems and Computer-Based Translation Tools*.
http://www.foreignword.com/Technology/art/Hutchins/hutchins99_1.htm
Consultada 23 de octubre de 2007.

Hutchins, W.J. 2000. *Early Years in Machine Translation: Memoirs and Biographies of Pioneers*. Amsterdam: John Benjamins.

Hutchins, W.J. 2001. "Machine Translation Over Fifty Years?". *Histoire, Epistemologie, Langage* 22, 1: 7-31.

Hutchins, W.J. 2005. *The History of Machine Translation in a Nutshell*
http://www.thocp.net/reference/machinetranslation/machinetranslation.html

Consultada 29 de octubre de 2007.
Meteo-System. http://en.wikipedia.org/wiki/Meteo_System
Última modificación 10 de febrero de 2007. Consultada 6 de noviembre de 2007.
OpenLogos. http://en.wikipedia.org/wiki/OpenLogos
Última modificación 21 de marzo de 2007. Consultada 6 de noviembre 2007.
Pinchuck, Isadore. 1977. *Scientific and Technical Translation*. London: Andre Deutsch.
Slocum, J. 1988. *Machine Translation Systems*. Cambridge University Press.
Systran. http://en.wikipedia.org/wiki/Systran
Última modificación 20 de octubre de 2007. Consultada 6 de noviembre 2007.
Traducción Automática
http://es.wikipedia.org/wiki/TraducciA3n_autoA¡tica
Última modificación 12 de octubre de 2007. Consultada 17 de octubre de 2007.