Traducción automática, corpus lingüísticos y desambiguación automática de los significados de las palabras[1]

Aquilino Sánchez
Universidad de Murcia
asanchez@um.es

1. Los inicios de la traducción automática (TA)

La traducción automática (TA) había constituido ya para algunos, años antes, un objetivo tan deseado como inalcanzable. Pero el pistoletazo de salida que llevó a iniciar los estudios sobre la posibilidad de automatizar la traducción lo dio W. Weaver (1949. Véase Hutchins, http://www.mt-archive.info/MTNI-22.pdf). Weaver era un investigador de la Fundación Rockfeller cuando hizo público un memorando que despertó el interés y la atención sobre la TA. En él, Weaver no solamente planteaba el tema de la automatización de la traducción, sino que incluía reflexiones de gran enjundia para esta incipiente disciplina. Nadie hablaba entonces, por ejemplo, del poder discriminante del contexto, o de su cometido en la definición del significado de las palabras. Weaver, sin embargo, ya hacía en su memorando observaciones tan capitales como las siguientes, sobre la desambiguación de los significados de las palabras polisémicas:

> If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words. […] But, if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then, if N is large enough one can unambiguously decide the meaning of the central word. […] The

[1] El presente estudio forma parte de las investigaciones que se realizaron dentro del Proyecto HUM2004-00080/FILO del Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica, del Ministerio de Educación y Ciencia.

> practical question is: "What minimum value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?

El autor manifiesta claramente que el cotexto constituye el marco decisivo para que las palabras *tomen* uno u otro significado cuando las posibilidades son múltiples. A tal fin, Weaver partía de algunas conclusiones basadas en el cálculo estadístico aplicado a la lengua (no en vano en 1949 publicó un libro sobre la teoría matemática de la comunicación, junto con Shannon). Es preciso añadir, sin embargo, que Weaver otorgaba entonces al cotexto un grado de simplicidad y poder discriminante que ha demostrado ser inadecuado en varios parámetros, amén de haber sido excesivamente optimista en cuanto a las posibilidades de ser controlado y utilizado como herramienta útil en la TA.

El escrito de Weaver despertó conciencias y atrajo el interés de muchos. Pero esta situación se prolongó poco en el tiempo. Los resultados no llegaban con la rapidez esperada y el sueño de la *máquina de traducción automática* fue desvaneciéndose con relativa prontitud. De nuevo volvió a reavivarse el interés por la TA. Tanto los Estados Unidos como Rusia necesitaban instrumentos que les permitiesen acceder con rapidez –a ser posible casi de manera instantánea– a la información del contrario. Además, la potencia de los ordenadores se incrementaba notablemente con el paso de los años. Los resultados de la TA seguían siendo toscos y precisaban de la revisión de traductores profesionales antes de ser dados por válidos. Pero no era menos cierto que a veces la traducción automática era suficiente para entender lo esencial de los artículos traducidos mediante ordenador. A pesar de ello, la década de los sesenta fue poco favorable para la TA. En 1966 se publicó el informe ALPAC (*Automatic Language Processing Advisory Committee.* Véase Pierce, Carroll, et al. 1966), a instancias del Gobierno Americano. Las conclusiones eran claramente pesimistas y poco alentadoras de cara al logro de cotas aceptables de eficacia traductora, y lo que es peor, no se vaticinaba tampoco un nivel de potencialidad o expectativas que invitasen al optimismo. Las ayudas estatales disminuyeron considerablemente y la actividad en torno a la TA se redujo notablemente. A consecuencia de ello, la investigación prácticamente desapareció de las universidades americanas (incluida la Universidad de Georgetown, que había patrocinado los primeros proyec-

tos). Por suerte, no ocurrió lo mismo en Canadá y en algunos otros países europeos, además de en Rusia.

En la década de los setenta surgió un sistema de TA que contribuyó decisivamente a mantener el interés sobre el tema: *Systran*, un programa basado en reglas (mucho más potente que el basado en un lexicón bilingüe). En 1976, la Unión Europea lo adquirió como instrumento de trabajo en sus oficinas y este hecho fue también decisivo para que la TA resurgiese de nuevo, incluso con ciertas perspectivas comerciales, como el fénix de sus cenizas. La necesidad de traducciones rápidas venía avalada por el incremento de los intercambios comerciales y la urgencia de que tales traducciones se pudiesen expandir a varias lenguas. Es interesante recordar que a finales del siglo XX, empresas tan conocidas como Fujitsu, Toshiba, NTT, Brother, Catena, Matsushita, Mitsubishi, Sharp, Sanyo, Hitachi, NEC, Panasonic, Kodensha, Nova u Oki competían con otras de mayor tradición en el área de la traducción, como era el caso –sobre todo– de IBM.

2. Nuevos paradigmas en la TA

Los diccionarios bilingües han sido los instrumentos tradicionalmente usados por los traductores humanos. De ahí que también fueran tomados desde el principio como recursos habituales en la TA. Pero los recursos utilizados por traductores humanos no son necesariamente igual de eficaces en su utilización por las máquinas. Los diccionarios constituyen un repertorio de significados y sus equivalencias, pero no incluyen ni los resortes ni las reglas que rigen su selección. Y eso es lo que precisan los ordenadores. De ahí que pronto se sintiera la necesidad de incluir reglas predeterminadas para tomar decisiones durante el proceso de traducción. El programa *Systran* fue pionero en esta nueva estrategia. Las reglas, no obstante, eran limitadas en su potencialidad. La emergente lingüística del corpus, junto con los estudios estadísticos aplicados al lenguaje, dieron lugar a otro procedimiento: el denominado *método estadístico*, basado en el análisis de grandes compilaciones textuales (corpus lingüísticos) y en el descubrimiento y aplicación de patrones recurrentes y sus equivalencias en la lengua a la que se pretende traducir. Dos factores enriquecen notablemente la TA e incrementan el optimismo de los investigadores: el acceso a ingentes canti-

dades de textos bilingües, equivalentes o paralelos, susceptibles de convertirse en modelos para futuras traducciones, y la potencia de los ordenadores, capaces ahora de procesar esa gran cantidad de información lingüística en pocos segundos. Sin lugar a duda, la conjunción de ambos instrumentos es decisiva en la consolidación y avance de la TA.

En este imparable proceso de búsqueda y perfeccionamiento de la TA se ha interpuesto continuamente la desazón de quienes temen que las máquinas sustituyan al traductor humano, privándole de la fuente de su sustento. Frente a estos temores, la reacción ha sido, invariablemente, infravalorar las traducciones realizadas por el ordenador. A pesar de ello, los buscadores de Internet (Altavista, Google, etc.) han percibido la apremiante necesidad de universalizar la información contenida en sus servidores y empiezan a ofrecer a los usuarios la posibilidad de traducir algunas de las páginas buscadas. Los resultados no siempre acaban *consolando* a los más pesimistas: a pesar de que en algunos casos la TA es aceptable y logra transmitir razonablemente bien el contenido del original, en otros casos los logros son manifiestamente inadecuados, e incluso incomprensibles. La realidad actual de la TA avala una razonable dosis de pesimismo, pero esto no es razón suficiente ni para desistir ni para echar las campanas al vuelo. Cuanto más se profundiza en el tema de la traducción, se hace más evidente que las herramientas y recursos disponibles todavía están muy lejos de colmar las necesidades existentes. Los diccionarios bilingües deben ser sustituidos por otro tipo de recursos, diseñados y elaborados específicamente para las máquinas; la complejidad del lenguaje natural se incrementa en el componente semántico –precisamente el más difícil de tratar automáticamente–, y las palabras, tradicionalmente consideradas como unidades léxicas del significado, no siempre pueden ser consideradas como tales, ya que la fijación de su significado depende o es activado por unidades co-textuales de mayor amplitud y extensión. Se está consolidando la idea de que el tratamiento automático del lenguaje mediante ordenador no ha creado aún ni los recursos ni las herramientas que necesita. Para que la TA pueda sustituir a los traductores humanos profesionales se ha de recorrer, pues, un largo camino. La complejidad del proceso de traducción aún no ha sido formalizada de manera satisfactoria.

Como ya apunté anteriormente, los repertorios léxicos bilingües se mostraron muy pronto insuficientes para responder a las necesidades de la TA. La gramática generativa contribuyó al desarrollo del *método basado en reglas*, tipo *Systran*. En los últimos años, las aportaciones más importantes en la TA han provenido (i) de los corpus lingüísticos y (ii) de la disciplina denominada *desambiguación automática de los sentidos de las palabras polisémicas* (*WSD*, *Word Sense Disambiguation*, en inglés).

2.1. Corpus lingüísticos: su utilidad

Los hablantes, especialmente los lingüistas, pueden ser muy eficaces en el análisis de los textos desde distintas perspectivas. Pero su capacidad para analizar textos está muy limitada en los aspectos cuantitativos. La lectura de un libro de 300 páginas puede ocuparnos uno o dos días. Un ordenador, sin embargo, es capaz de leer ese mismo libro en cuestión de segundos. Y no sólo eso: es capaz de almacenar cada una de esas palabras leídas, ordenarlas todas ellas por orden alfabético, o contarlas y luego ordenarlas por orden de frecuencia, etc., etc. En cuanto a la velocidad de lectura y procesamiento de las palabras, un ordenador es tan superior a nuestra capacidad que la comparación resultaría odiosa… para el ordenador.

Kaeding (Sánchez 1995, citando a Atkins y Zampolli 1994: 21) recopiló un corpus de 11 millones de palabras en 1897 *(Häufigkeitswörterbuch der deutschen Sprache)*. Pero murió antes de alcanzar los resultados que habían motivado tal recopilación: contar y analizar la frecuencia de vocales y consonantes y sus patrones de organización. Un ordenador actual habría hecho ese mismo trabajo en pocos segundos. La *palabra digital* ha abierto una nueva dimensión en el procesamiento y manipulación del lenguaje humano. Y sobre todo, nos ha permitido analizar la producción lingüística de miles o millones de hablantes, tanto del lenguaje oral como del escrito. El acceso a tales volúmenes de palabras, aisladamente o en sus correspondientes contextos (discurso), ha propiciado un cambio radical en los estudios lingüísticos. Al igual que en otras áreas de la lingüística, la TA no ha podido sustraerse a las ventajas de la lingüística basada en corpus.

Las palabras en cuanto unidades formales del lenguaje, como secuencias de letras, limitadas y bien definidas, constituyen elementos

perfectamente adecuados para su tratamiento mediante ordenadores. Cada palabra puede ser identificada fácilmente, puesto que está separada de la precedente y de la siguiente por un espacio en blanco; puede ser objeto de cuenteo, puede ser ordenada en cualquier secuencia y a una velocidad impensable para el ser humano. Los datos así obtenidos están perfectamente objetivados, pueden ser sistematizados, agrupados de acuerdo con determinados esquemas o patrones, y con información numérica de distinta índole. Pero a la hora de llevar estos resultados al campo de la traducción, no podemos perder de vista que las palabras, aisladamente, fuera del contexto comunicativo en que se dan, son en buena parte *entidades muertas*. Las palabras sólo cobran vida en el discurso en que son utilizadas. Esta realidad se pone más claramente de manifiesto en los términos polisémicos. La palabra inglesa *saving* cobra significado si decimos *Saving money*. Y el contraste se revela de inmediato si la usamos en otro contexto, como en *Saving a girl*. Con la simple adición de una *s* percibimos también otro contraste significativo: *My savings are in the bank.*

En la medida en que las palabras se enriquecen con el significado (en la medida en que cobran *vida semántica*), incrementan su complejidad y pierden las grandes ventajas que tienen las formas puras para ser procesadas mediante ordenador. De momento, el componente semántico en su plenitud apenas si está al alcance de un ordenador. La culpa de este fracaso no hay que atribuirla al computador, sino a los seres humanos que los hemos creado. Si el ordenador supiese cuándo *saving* significa *salvar* o cuándo significa *ahorrar* o *ahorro*, el problema estaría solucionado. Para ello sería necesario que sus creadores –el hombre– le suministrasen esta información. Curiosamente, los hablantes sabemos cuándo hay que elegir cada significado, pero no somos capaces de analizar el proceso con la suficiente precisión y detalle para identificar todos los elementos en que nos basamos para hacerlo.

Hemos avanzado algo: no bastan los glosarios bilingües, ni las reglas formales conocidas y expresadas en las gramáticas. Y en los últimos años se ha empezado a utilizar otro recurso: el análisis de patrones recurrentes en el lenguaje, o en la traducción de una lengua a otra. Hasta ahora era difícil tener acceso a esa información, ya que la sola introspección individual o la observación limitada que un investigador puede llevar a cabo sobre la totalidad del lenguaje es insuficiente para la obtención de datos fiables. Aquí es donde los corpus lingüísticos

vienen a llenar un hueco. El procesamiento de grandes cantidades de muestras lingüísticas (palabras tal cual aparecen en la comunicación real de los hablantes) implica grandes ventajas: no solamente podemos recabar información sobre las formas (su morfología y sintaxis), sino también sobre la relación que se da entre el significado y la disposición y uso de esas formas en el discurso: su posición en la frase, la frecuencia de su presencia en determinados contextos, quiénes las usan, en qué tipo de discurso aparecen, en qué combinaciones se utilizan, qué patrones son recurrentes, etc.

La traducción puede beneficiarse de los corpus monolingües y bilingües. Los primeros porque nos pueden dar información sobre el léxico o los conjuntos léxicos que dan origen a los significados en una lengua determinada. Los segundos porque nos permitirán relacionar las correspondencias entre los grupos léxicos, las palabras y sus significados en cada una de las lenguas y establecer, en consecuencia, comparaciones, semejanzas y equivalencias. Si la fijación de significados la determina el uso lingüístico, la detección de ese uso, tanto en lo relativo a formas como a patrones, y su relación con las equivalencias en otra lengua es el mejor aval para crear nexos automáticos útiles en la traducción. En esta premisa se asienta la bondad de los corpus para la TA. Los corpus son una base excelente para investigar el uso, y especialmente el uso recurrente, detectable por su índice de frecuencia.

La TA basada en ejemplos surgió como método previsiblemente capaz de superar y mejorar el método de TA basado en reglas. Las diferencias entre ambos métodos son evidentes: las reglas generan modelos partiendo de enunciados generales y produciendo ejemplos (*top-down*), mientras que el método basado en ejemplos actúa en sentido contrario: los modelos resultantes se producen tras la observación de ejemplos reales detectados en el uso (*bottom-up*). El modelo *de arriba-abajo* es ideal y eficaz cuando las reglas formuladas son representativas de la realidad y en la medida en que se ajustan a ella. Pero no es aplicable cuando las reglas o no existen o no pueden ser formuladas con precisión. En cambio el modelo de *abajo-arriba*, a pesar de que no alcance la perfección en su potencialidad para ser proyectado a la totalidad de los casos posibles, sí que alcanza un alto grado de fiabilidad siempre que la frecuencia en que se fundamenta sea la adecuada. Este proceso no es ajeno para quien aprende lenguas: podría aventurarse que las reglas que los hablantes inducen o construyen durante el

período de adquisición de la lengua materna no tienen otra fuente que el uso detectado por el aprendiz, y especialmente, el uso recurrente. Un ejemplo ilustrativo podría ser la asociación entre la terminación *–o* y el género masculino, o la terminación *–a* y el género femenino: la repetición sistemática de este comportamiento lingüístico lleva a los hablantes nativos de español y a los estudiantes extranjeros a aplicar sistemáticamente el patrón detectado en la práctica. El problema de los alumnos extranjeros surge cuando este patrón no se aplica, como es el caso de *problema*. El hecho de que los aprendices tiendan a decir *la problema* en vez de *el problema* pone de manifiesto que no conocen expresamente las excepciones a tal patrón o regla; de ahí que no sean capaces de rectificarlo en ciertos casos minoritarios, o, lo que es lo mismo, sigan aplicándolo con carácter general.

Dos son las modalidades de corpus multilingües útiles para la TA: los corpus *comparables* y los corpus *paralelos* (del inglés, *comparable* y *parallel corpora*) (para mayor información sobre los distintos tipos de corpus puede consultarse, entre otros, McEnery y Wilson 1996). En los corpus equivalentes o comparables, el tamaño de la recopilación textual es semejante, y los textos se asemejan en cuanto al contenido, si bien el grado de semejanza puede ser variable. Lo más típico de este tipo de textos bilingües es que estén originados por la misma causa, es decir, por la misma información. Por ejemplo, una noticia relatada en dos lenguas diferentes. Ha de tenerse en cuenta que la redacción de cada texto no es predecible, ni se ajusta a las mismas reglas o normas, desde cualquier punto de vista que se considere el tema. Sólo el punto de partida comparte idéntica base: la noticia o el tema son los mismos. La subjetividad del narrador o escritor estará siempre presente, las estructuras sintácticas empleadas pueden divergir notablemente y el vocabulario seleccionado es posible que sea bastante dispar. Por razones similares, pueden darse notables diferencias en la abundancia y variedad léxica, e incluso en la cantidad de términos utilizados para la narración. En consecuencia, un *corpus equivalente* es útil para estudios o análisis en contextos amplios, o para estudios de tipo pragmático en los que se analiza cómo un hecho es contado en dos o más lenguas. Véase cómo la misma noticia puede dar origen a dos relatos bien dispares: tanto la perspectiva desde la que se enfoca la noticia como los hechos a los que se alude en la misma se distancian tanto en cuanto al léxico utilizado que la comparación se debe reducir casi a constatar las diferencias:

Tabla 1. *The New York Times* (USA) – *El Mundo* (España)

In a national address on state television, Mr. Musharraf said that Pakistan had been in extreme danger and that imposing the emergency had been "unavoidable." Since then, he said, the situation had largely improved, armed militants had been pushed back in the northwest of the country, and free and fair elections would be held across the country on Jan. 8.	El *estado de excepción vigente* en Pakistán desde el pasado 3 de noviembre ha quedado anulado por orden del presidente del país, Pervez Musharraf, según ha informado el fiscal general del Estado.
"Today I am feeling very happy that all the promises that I have made to the people, to the country, have been fulfilled," he said. The removal of the state of emergency restores fundamental rights like the right of assembly and freedom of movement three weeks before parliamentary elections, and it would ensure that elections are free and fair, said the acting law minister, Afzal Haider.	En un comunicado, el fiscal general Malik Qayyum ha anunciado el restablecimiento de la Constitución paquistaní y el fin de la Orden Constitucional Provisional por la cual Musharraf ha decretado el estado de excepción. Qayyum, cercano colaborador de Musharraf, ha asegurado que el presidente ha cumplido sus compromisos y adelantó que las *elecciones legislativas del próximo 8 de enero podrán celebrarse ahora con normalidad.*
Mr. Musharraf also took the oaths of 14 new Supreme Court judges Saturday afternoon, permanently replacing the Supreme Court he dismissed on Nov. 3.	La restauración de la Constitución de 1973 incorpora las enmiendas realizadas por Musharraf para blindar la validez de la excepción y asegurarse de que su reforma del Tribunal Supremo no es revocada.
In repealing the state of emergency, which many here described as de facto martial law, Mr. Musharraf has completed a number of steps demanded by his critics at home and abroad, and by the Bush administration, to return the country to the path to democracy. On Nov. 28, he resigned his military post of chief of army staff, ending eight years of military rule.	Las últimas enmiendas, aprobadas el viernes, pretenden dar continuidad a los jueces del Supremo que juraron el cargo tras la declaración de la excepción, así como el cese definitivo de los anteriores. Los *nuevos jueces*, más afines a Musharraf que los anteriores, *deberán jurar de nuevo sus cargos ante la Constitución*, ya que sólo lo hicieron ante la Orden Provisional una vez que el presidente declaró el estado de excepción.

Naturalmente, no todos los relatos de la misma noticia presentan diferencias tan notables. Pero cabe esperar que habitualmente sean bastante distantes, al menos en cuanto al léxico utilizado. De ahí que en la TA sean de mayor utilidad los corpus *paralelos* o *en paralelo*.

Los corpus *paralelos* están constituidos por una lengua A (lengua original) traducida a otra lengua B (lengua meta), o incluso a otras lenguas (C, D, etc.), si se trata de corpus multilingües. El hecho de que lo que origina la traducción sea un texto ya definido cambia radicalmente los resultados que cabe esperar de la traducción. De ahí que su utilidad para la traducción en general o para la TA sea sustancialmente distinta, si la comparamos con los corpus equivalentes.

2.2. Corpus paralelos y TA

El uso de los corpus como recurso útil en la traducción es reciente y en este caso sí proviene del mundo académico. Es un hecho reseñable, porque hasta no hace muchos años, los estudios de traducción ni siquiera eran tomados en consideración por los centros universitarios. Incluso cuando la lexicología ya ocupaba un puesto de relieve dentro de los estudios lingüísticos y figuraba como disciplina obligatoria en los currículos, la lexicografía y la traducción eran responsabilidad de las empresas editoriales o de autores individuales. La historia de los diccionarios da fe de ello. La lexicografía no merecía la consideración del mundo académico porque era una disciplina *práctica*. La conexión entre lexicología y lexicografía es evidente, pero solamente en los últimos años ha empezado a ser tratada en profundidad. Y la evidencia extraída de los corpus ha contribuido a ello. Son muchos los estudios que ponen de manifiesto la interdependencia o dependencia léxica (Almela 2006; Hoey 1991, 2005; Sinclair 1991, 2004), y por lo tanto la importancia del contexto para definir el significado de cada palabra (Cantos y Sánchez 2001). En realidad, el repertorio de significados de los diccionarios cobra validez en el discurso; aisladamente está incompleto, comunicativamente es sólo material de referencia. El análisis basado en corpus monolingües ha reafirmado todos estos extremos. Solo faltaba ampliar los estudios a los corpus bilingües. En efecto, si la dependencia léxica es un hecho en la utilización de las palabras, y siendo así que tal dependencia léxica no es igual en todas las lenguas,

se hace necesario identificar tales dependencias en cada idioma para poder establecer comparaciones, o para ofrecer equivalencias fiables.

Baker (1999) afirma que los corpus son útiles en la traducción, especialmente –dice– en la enseñanza de la traducción. La afirmación no es novedosa, pero contribuye a impulsar el aprovechamiento de recursos actualmente accesibles y, sobre todo, útiles. Laviosa (1997, 1998) añade que los corpus bilingües aportan muestras y hechos que permiten elaborar o formular *leyes de comportamiento traductológico* (*laws of translational behaviour*), es decir, reglas de equivalencias basadas en cálculos probabilísticos. Quizás debería hablarse, en estos casos, no tanto de reglas cuanto de modelos basados en ejemplos, según lo apuntado anteriormente.

Fue un corpus elaborado en la década de los setenta y ochenta (*The Hansard Corpus*) el que se tomó como modelo de corpus paralelo. Este corpus consta de textos del parlamento canadiense alineados en francés y en inglés. Si en un principio el análisis de estos materiales se hacía tedioso y agobiante, el advenimiento de nuevas herramientas de programación y la potencia de los ordenadores han facilitado enormemente el trabajo. Dos textos alineados frase por frase, o párrafo por párrafo son útiles para estudiar las equivalencias puntuales en una y otra lengua y extraer consecuencias, generalmente limitadas en su proyección. Pero si a este análisis puntual le añadimos el análisis del conjunto, como lo puede hacer un ordenador, computando con exactitud cuándo, cuántas veces y en qué contexto un término tiene una determinada equivalencia en la otra lengua, o cuándo, cuántas veces y en qué contexto un grupo léxico o patrón tiene una determinada traducción en la lengua meta, en tal caso es posible generar una *regla* probabilística.

Los corpus paralelos, para que resulten más eficaces o plenamente satisfactorios para los fines de TA, deben de estar perfectamente alineados, a ser posible frase por frase. Solo así podemos establecer comparaciones e identificar equivalencias. Si prestamos atención a los ejemplos de la tabla siguiente, extraídos de un corpus paralelo, descubriremos rápidamente que la frase en inglés *It may seem strange that I should think*... ha sido traducida al español como *Pudiera parecer extraño que yo haya creído*... O que *It is vital that we establish*... se corresponde con … *es vital establecer*…

Tabla 2.

It may seem strange that I should think it necessary to give such prominence to this element in the case of an author whom I have called a genius and a prophet.	*Pudiera parecer extraño que yo haya creído necesario dar tanta importancia a este elemento tratándose de un autor a quien he calificado de genio y de profeta.*
It is vital at this point that we establish diplomatic relations and therefore a dialogue with the current Kabul authorities.	*En este momento es vital establecer relaciones diplomáticas, y por tanto diálogo, con las autoridades actuales de Kabul.*

La alineación de textos permite detectar de inmediato los fallos o aciertos de cualquier traductor automático. El texto anterior, en una y otra dirección, es traducido así por el programa *Systran*:

Tabla 3.

Systran: Inglés >>>	>>>Español
It may seem strange that I should think it necessary to give such prominence to this element in the case of an author whom I have called a genius and a prophet.	Puede parecerse extraño que debo pensarlo necesario para dar tal prominencia a este elemento en el caso de un autor a que he llamado un genio y un profeta.
It is vital at this point that we establish diplomatic relations and therefore a dialogue with the current Kabul authorities.	Es vital a este punto que establecemos relaciones diplomáticas y por lo tanto un diálogo con las autoridades actuales de Kabul.
Systran: Español>>>	>>>Inglés
Pudiera parecer extraño que yo haya creído necesario dar tanta importancia a este elemento tratándose de un autor a quien he calificado de genio y de profeta.	*It could seem strange that I have believed necessary to give as much importance to this element being an author to whom I have described as genius and prophet.*
En este momento es vital establecer relaciones diplomáticas, y por tanto diálogo, con las autoridades actuales de Kabul.	*At this moment he is vital to establish diplomatic relations, and therefore dialogue, with the present authorities of Kabul.*

Se han subrayado las palabras o frases cuya traducción presenta algún problema. Y llama la atención el hecho de que tanto en la traducción inglés-español como en la de español-inglés los errores –de mayor o menor entidad– tienden a aparecer en los mismos puntos (*parecerse, debo pensarlo necesario, un*, etc.; *I have believed necessary, as much, being an uthor to whom, he*…).

El mismo análisis sobre los resultados de otro programa de TA, *Global Power*, presenta problemas muy similares, como puede apreciarse en el cuadro siguiente:

Tabla 4.

Global Power T.: Inglés >>>	>>>Español
It may seem strange that I should think it necessary to give such prominence to this element in the case of an author whom I have called a genius and a prophet.	Puede parecer extraño que yo deba pensarlo necesario dar la tal prominencia a este elemento en el caso de un autor quien yo he llamado a un genio y un profeta.
It is vital at this point that we establish diplomatic relations and therefore a dialogue with the current Kabul authorities.	Es a estas alturas vital que nosotros establecemos las relaciones diplomáticas y por consiguiente un diálogo con las autoridades de Kabul actuales.
Global Power T.: Español>>>	>>>Inglés
Pudiera parecer extraño que yo haya creído necesario dar tanta importancia a este elemento tratándose de un autor a quien he calificado de genio y de profeta.	*It could seem strange that I have believed necessary to give so much importance to this element being an author to who I have described as genius and of prophet.*
En este momento es vital establecer relaciones diplomáticas, y por tanto diálogo, con las autoridades actuales de Kabul.	*At this time it is vital to establish diplomatic relationships, and therefore dialogue, with the current authorities of Kabul.*

La elaboración de un corpus paralelo con datos como estos permitiría identificar los puntos problemáticos de la traducción. En los párrafos anteriores, constatamos que algunos tiempos y formas de los verbos en español no se corresponden con sus equivalentes en inglés. Puede

observarse que el uso del subjuntivo en español se sustituye por *modal + infinitivo* en inglés (*should think*). La constatación de este hecho en centenares, miles o decenas de miles de oraciones similares avalarían fiablemente el establecimiento de un patrón de traducción que reflejase tales equivalencias. La observación del comportamiento de algunos programas de TA en esos mismos casos incrementaría la fiabilidad del diagnóstico y de la solución que podría darse al mismo.

La alineación de un corpus bilingüe es un trabajo ímprobo y *delicado*. Cada oración de la lengua original no siempre se traduce por otra oración en la lengua meta. A veces dos oraciones se reducen a una, o una oración se desglosa en dos. Puesto que el objetivo es detectar con exactitud qué palabra o qué grupo de palabras en una lengua corresponde a cada palabra o grupo de palabras en la otra lengua, cualquier desvío meramente formal de una lengua respecto a la otra, resultará problemático e incrementará las dificultades de identificación, especialmente por parte de un sistema automático. Se han diseñado algunos alineadores automáticos para facilitar la tarea de comparación (véase, por ejemplo, Gelbukh and Sidorov 2006; Kay and Roscheisen 1993; McEnery and Oakes 1996). Pero los resultados aún deben perfeccionarse, ante las innumerables dificultades que concurren en el caso. No obstante, la investigación sobre el tema acaba de empezar y no es aventurado ser optimista en el logro de mejores resultados. Algo similar ocurrió con la etiquetación morfológica automática; actualmente, sin embargo, ya se logran índices de éxito en torno al 95%. De otra parte, la alineación automática de textos bilingües quizás no pueda separarse totalmente del proceso de TA; en tal caso, los avances en un campo irán de la mano de los avances en el otro.

En los últimos años, los corpus paralelos han sido también utilizados como ayuda apara la desambiguación automática de sentidos en las palabras polisémicas (Chan et al., 2004; Tufis et al. 2005). El logro de este objetivo requiere la alineación correcta de las palabras y frases en cada lengua y la garantía previa de que las equivalencias de la traducción son correctas. Se parte, además, de la premisa de que cada una de las acepciones de las palabras polisémicas suelen traducirse por una palabra diferente en la lengua meta (la palabra *wood*, en inglés, puede tener los equivalentes de *bosque* o *madera*, en español, siendo cada una de las palabras indicativas de dos sentidos diferentes). Dadas estas condiciones, si una palabra –o cada una de las acepciones

de una palabra– en la lengua A es traducida con una palabra específica en la lengua B, esta palabra de la lengua B puede tomarse como diferenciadora del significado o acepción que corresponde a la palabra polisémica de la lengua A. A partir de ahí se podrá concluir, pues, que la palabra A queda desambiguada en el sentido especificado por la palabra de la lengua a la que se ha traducido. El recurso a los corpus paralelos podría usarse por tanto como una ayuda para establecer los diferentes sentidos que un término puede contener (véase el apartado V, con información más detallada sobre la desambiguación automática de sentidos y sus implicaciones).

3. Traducción por profesionales y TA

Una de las diferencias clave entre el traductor humano y la TA reside en que mientras el primero puede analizar, reflexionar, decidir, volver a revisar, corregir, etc., en un proceso que puede prolongarse tanto tiempo cuanto quiera el mismo traductor, la TA carece de tal flexibilidad. Para que la TA se equiparase al traductor humano, sería necesario que el proceso del traductor humano fuese totalmente previsible y fuese enseñado a la máquina como tal. Pero por desgracia, una buena parte del producto del traductor humano o no es totalmente previsible o encierra ambigüedad (es decir, se enfrenta a más de una opción). En la medida en que exista imprevisión y ambigüedad, no cabe esperar que la TA logre las metas que los traductores humanos son capaces de alcanzar. Lo que es imprevisible no puede ser formalizado. Pero sí es posible prever una gran parte del proceso traductor mediante el análisis de cómo los elementos o ciertos conjuntos de elementos de una lengua son traducidos a otra. Dado que los corpus paralelos son susceptibles de aportar gran cantidad de información sobre las equivalencias entre dos lenguas, estos instrumentos constituyen recursos insustituibles y compañeros necesarios de la TA.

Los corpus bilingües paralelos han sido y siguen siendo utilizados en dos modelos de TA: el modelo estadístico y el modelo basado en ejemplos. Ambos modelos comparten el mismo punto de partida: precisan de una voluminosa base de datos como soporte para la elaboración de hipótesis traductológicamente válidas.

En realidad, el *modelo estadístico* lo inició el mismo Weaver en su obra conjunta con Shannon (Shannon and Weaver 1949). La decisión para optar por una traducción concreta se fundamenta en el mayor grado de probabilidad de esa traducción frente a otra u otras. De manera que si la traducción B es la más probable, entre las analizadas, a partir del texto A, en tal caso la máquina opta por la opción B. Simplificando el tema –y corriendo el riesgo de caer en excesivas simplificaciones–, la probabilidad en la que se basa el método estadístico puede formalizarse así: $p(e \mid f)$, es decir, la cadena *e* en la lengua origen se corresponde con la cadena *f*, que es la traducción en otra lengua (por ejemplo el español). Naturalmente para llegar a una conclusión de este tipo, el sistema debe comprobar que todas las cadenas *e* en la lengua origen producen unos determinados resultados (varias o muchas cadenas *f*), entre los cuales la cadena *f* seleccionada es la que cuenta con mayor probabilidad de ser la más adecuada. Adviértase que el modelo estadístico se puede aplicar tanto a palabras como a sintagmas u oraciones. De hecho, las primeras aplicaciones se llevaron a cabo con palabras y posteriormente se ha ido ampliando la cadena de elementos o formas tomados en consideración para el cálculo de la probabilidad.

El *modelo basado en ejemplos* también se fundamenta en la observación de las equivalencias observadas en textos originales y sus traducciones (corpus paralelos), pero intenta imitar el método seguido por los humanos en procesos similares. De ahí que algunos afirmen que es un método fundamentado en la *analogía* (Nagao 1984), es decir, transfiriendo la información observada en un campo determinado y concreto (fuente) a otro campo también concreto (objetivo). Es un proceso cognitivo, pero no un proceso inductivo o deductivo. En el caso de la traducción, el campo concreto que actúa como fuente es la observación detectada en la lengua origen, mientras que el campo concreto que constituye el objetivo es la lengua a la cual se traduce.

El modelo por analogía se inspira en la manera como el ser humano *traduce* de una lengua a otra cuando inicia el aprendizaje de un idioma extranjero. Según Nagao (1984),

> (1) Man does not translate a simple sentence by doing deep linguistic analysis, rather,
> (2) Man does the translation, first, by properly decomposing an input sentence into certain fragmental phrases (very often, into

case frame units), then, by translating these fragmental phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference, which is illustrated above.

Un modelo tal requiere de un largo y cuidadoso entrenamiento, puesto que primero es preciso proceder a la fragmentación de oraciones y sintagmas con el fin de buscar cada caso en el que la analogía de la traducción sea aplicable. Se trataría, por ejemplo, de comparar dos frases como las siguientes, traducidas del inglés al español:

(1) How old is **your daughter**? ¿Qué edad tiene **tu hija**?
(2) How old is **my cousin**? ¿Qué edad tiene **mi primo**?

La fragmentación de estas frases podría plantearse así:

(3) *How old is* **X**? corresponde a *¿Qué edad tiene **X**?*
(4) *Your daughter* corresponde a *tu hija*
(5) *My cousin* corresponde a *mi primo*

Con estos ejemplos la máquina habría aprendido tres unidades de traducción, pero sobre esa base podría traducir automáticamente cualquier ejemplo en el que *X* esté inserto en una estructura similar, como sería, *How old is **the father of your husband**?, How old is **the car you bought***?, etc. Teniendo en cuenta, eso sí, que los fragmentos pueden ser también complejos y precisar de una sub-fragmentación reconocible por el sistema. La máquina debe ser capaz de fragmentar cada frase en componentes parecidos a los de (3), (4) o (5), o en otros similares previamente definidos. Y aquí es donde los corpus paralelos podrían aportar una valiosa ayuda, proporcionando al ordenador materiales suficientes para ser entrenado en las oraciones pertinentes y en los grupos léxicos o sintagmas en los que pueden ser fragmentados, con su correspondiente equivalencia en la lengua meta.

De lo aquí expuesto cabe concluir que la TA dista aún mucho de haber logrado un estadio satisfactorio. También es evidente que la TA no puede prescindir todavía de la supervisión del traductor profesio-

nal. Y sobre todo que el uso de los corpus paralelos, en cualquiera de los métodos adoptados, es un requisito indispensable, pero teniendo en cuenta que los datos obtenidos del análisis o simple uso de tales recursos precisarán de la intervención humana para ser validados. La complejidad del lenguaje es tal que sería ilusorio caer en simplificaciones. Es más, la complejidad existente en el sistema lingüístico de comunicación no debe considerarse como algo que llega a un tope en el cual se estabiliza. Más bien al contrario, la complejidad del lenguaje se presenta como un constructo en continuo crecimiento, lo cual hace presagiar que la TA será también un objetivo al cual nos iremos hacercando más y más, pero que nunca se alcanzará en su totalidad.

4. Polisemia y desambiguación automática de los sentidos de las palabras.

Los problemas de la TA están íntimamente ligados al alto nivel de polisemia léxica y consecuente ambigüedad a que este hecho da lugar. La ambigüedad es un problema de las palabras (Almela 2006; Sánchez, Cantos y Almela, 2007; Sinclair 1991, 2004), y no de los textos. El discurso lleva consigo las claves necesarias para desambiguar los términos semánticamente polivalentes. De ahí que el discurso no suela incluir ambigüedad, hecho un tanto insólito a primera vista, ya que si tomamos como punto de referencia los más de 15 significados que pueden tener lemas como *mano/hand, ir/go, echar, get*, etc., estaríamos abocados a ser más bien pesimistas.

La TA se apoyó en sus inicios en repertorios léxicos bilingües, con la información habitual disponible en tales herramientas. Pero en contra de las expectativas iniciales, la multiplicidad de opciones ofrecidas en la traducción de muchos términos puso de manifiesto la dificultad de trasladar los significados pertinentes –y sólo los pertinentes– a otra lengua. Es posible argumentar que el proceso de la traducción no descansa sobre las palabras aisladas, sino sobre unidades más amplias, como son la oración, el párrafo, o el discurso en general. La TA es ahora más consciente de ello y tiende, como vimos en páginas anteriores, a tener en cuenta esta realidad para incrementar la calidad de la traducción. Pero en cualquier caso, el discurso captado por los hablantes pierde la ambigüedad inherente a muchas de las palabras que lo in-

tegran precisamente porque los hablantes son capaces de interpretar con rapidez las claves desambiguadoras que el texto lleva consigo. Esta capacidad de interpretar las claves desambiguadoras del discurso es la que aún no han logrado las máquinas usadas en la TA. El método por analogía (o basado en ejemplos) puede eliminar algunos problemas de ambigüedad debido a la potencialidad que tiene de captar el significado exacto de fragmentos del discurso. De igual manera, el método basado en la estadística es también capaz de solucionar positivamente problemas similares del texto en unidades superiores a la palabra.

Las claves para fijar el significado léxico pueden estar situadas en varios niveles: a veces en el nivel oracional, a veces en el nivel sintagmático, y muy a menudo en el nivel léxico. La TA deberá, pues, atenerse a esta realidad. Por lo general, la decisión automática se toma tras responder a preguntas como las siguientes: si nos enfrentamos a una palabra A (preferentemente arropada por un co-texto), y si esta palabra es precedida o seguida por la palabra B (+ C, D, etc., según los casos), ¿hay alguna posibilidad de que las equivalencias en el texto meta (texto traducido) sean S, X o Z? Esta condición puede plantearse tantas veces cuantas sea necesario. Pues bien, para que esta pregunta la pueda hacer el sistema y para que pueda responderla adecuadamente, es necesario que disponga de toda la información implicada en tal decisión. En el campo del léxico, la información requerida implicaría, como mínimo:

1. Tener registrados todos los posibles sentidos de una palabra en la lengua A y sus equivalentes en la Lengua B.
2. Tener registradas las condiciones en que cada sentido se aplica, primero en la lengua A y luego en su relación con la lengua B (no siempre hay correspondencia plena entre ambas, como bien saben los traductores humanos).
3. Conocer los co-textos y contextos en que cada significado se materializa y las equivalencias de tales co-textos y contextos en la lengua meta (no siempre se dan equivalencias entre dos sistemas lingüísticos en este ámbito).

Estas condiciones son más complejas de lo que aparentemente pudieran sugerir. Los diccionarios actuales suelen cumplir razonable-

mente bien la condición 1. Pero en cuanto a las condiciones 2 y 3 la mayor parte del camino está por recorrer. Realmente lo que se necesita es un *DNI* de cada palabra. Los identificadores ofrecidos por los diccionarios son útiles pero incompletos para los fines de la TA o de la desambiguación automática. Los diccionarios, de hecho, han sido tradicionalmente desarrollados para ser tomados como referencia por hablantes de la lengua, es decir, por conocedores de los co-textos, los contextos y demás condiciones que rigen el discurso. Los ordenadores, en cambio, carecen de conocimientos sobre esos extremos. De ahí, pues, la necesidad de elaborar nuevas herramientas adaptadas a las necesidades de las máquinas. Los recursos propiciados por los corpus constituyen una ayuda excelente para el logro de tales fines.

La desambiguación de los diferentes significados de las palabras polisémicas está en la raíz de la gestión automática de los significados. Y la desambiguación automática de significados (DAS) necesita de una comprensión adecuada de su organización en las palabras para proceder a una correcta desambiguación cuando sea necesario. Los diccionarios tradicionales han contribuido sin lugar a duda a que los hablantes consideren los significados de cada palabra como un aglomerado de unidades significativas ordenadas secuencialmente, tal cual suelen aparecer en las obras lexicográficas. Los lingüistas saben, sin embargo, que la organización de los significados en las palabras es a menudo poco transparente, más bien borrosa y difícil de captar incluso para el analista. De entrada, los diferentes rasgos semánticos mediante los cuales suele describirse el significado léxico están íntimamente relacionados entre sí, en relación jerárquica, pero no necesariamente lineal y en un solo plano, sino más bien en diferentes planos o dimensiones. El concepto de *red volumétrica* describe la organización semántica de las palabras mucho mejor que el concepto de *malla extendida sobre un solo plano*.

El tipo de definición lexicográfica tradicional está fundamentado en el llamado *genus et differentiae*. Este modelo definitorio asume que el significado de las palabras, consideradas como unidades léxicas, se descompone en rasgos semánticos. A su vez, la técnica definitoria empieza especificando el género o la clase a que pertenece la cosa definida, para detallar a continuación las características propias de esa clase; de esta manera se diferencia una clase de otras y se distinguen entre sí los objetos que pertenecen a la misma clase. Así, la definición de ti-

gre, puede enunciarse como *animal mamífero* (clase) y completarse con rasgos como *felino, carnívoro, fiero, parecido a un gato de gran tamaño, de piel roja amarillenta o anaranjada con rayas negras en el lomo y en la cola, y de color blanco en el vientre. Tiene una gran fuerza muscular, gran agilidad y alta velocidad en la carrera.* La definición es adecuada para distinguir a un tigre de un león (éste no tiene rayas en la piel), o de una vaca (ésta no es carnívora), o de una mesa (la mesa no es animal). Los diccionarios están, pues, organizados y elaborados para poner de manifiesto los contrastes significativos de las palabras como elementos autónomos y unidades aisladas. Queda luego a cargo de los hablantes o usuarios de la lengua la responsabilidad y la habilidad necesarias para identificar en cada caso el significado concreto que se pretende transmitir en la comunicación real, es decir, cuando las palabras se insertan en el discurso y en él dejan de ser polisémicas. Para llevar a cabo eficazmente este cometido, los diccionarios apenas ofrecen ayuda específica. La DAS carece de las herramientas o de la información necesaria para realizar esta labor. Los humanos llevamos a cabo esta tarea de manera inconsciente, pero eficaz. ¿En qué nos basamos para ello? Considérese esta oración:

(6) Al frente de las tropas enemigas mandaba *un tigre* con alta graduación militar.

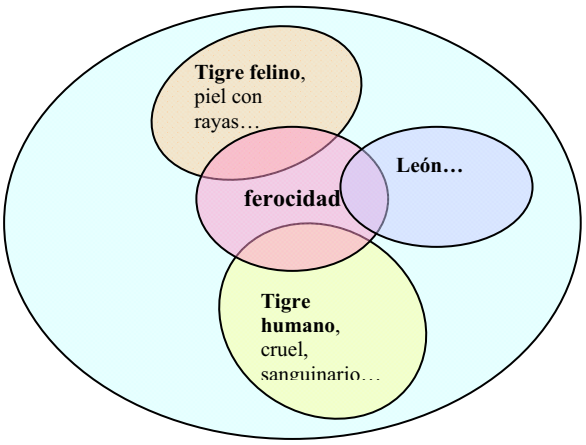
Además del significado de *tigre* como *animal felino*, *etc*., los diccionarios incluyen otro significado en su repertorio, aplicable al detectado en esta frase:

(7) Persona que es cruel y agresiva.
Persona cruel y sanguinaria.

Existe una clara relación entre esta definición de *tigre* y la más habitual (de *animal felino*…). El segundo significado de *tigre* es el denominado tradicionalmente *figurado*. Se entiende por tal que el significado se ha generado mediante un proceso por analogía respecto al original, el *tigre real*: el término se aplica también a una persona cuando ésta tiene algunas de las características propias del *tigre real*. Estas características, según los diccionarios, son ‘crueldad, agresividad y pla-

cer en el derramamiento de sangre (sanguinario)'. Ninguno de estos rasgos semánticos se menciona exactamente en la definición primera. Y realmente no hay razón para considerar que un tigre sea cruel o sanguinario: mata para sobrevivir, no necesariamente por placer. Uno de los rasgos del tigre real (*ferocidad*) lleva a aplicar este mismo nombre a una persona, pero *enriqueciéndolo* con características o matices propios de la perspectiva humana: la *ferocidad* del ataque lleva generalmente a la crueldad y un ataque feroz suele ser *sangriento*. Así pues, la segunda definición implica una transposición o *traslado* del significado primero a otro plano o dimensión diferente: el ámbito del comportamiento humano. Gráficamente, podría representarse esta conjunción de significados de la siguiente manera:

Figura 1.



Los dos *tigres* comparten plenamente el carácter de *animal y mamífero* (clase a la que pertenecen). Pero en cuanto a rasgos identificadores, solo comparten el de *ferocidad*, no el de *felino*, ni *crueldad*, por ejemplo. Lo que une a ambos *tigres* es el rasgo de *ferocidad* que conllevan, rasgo que también caracteriza al *león*, por ejemplo, y por eso este término también puede aplicarse al ser humano (piénsese en *el león de Judá*, por referencia al rey David). La palabra *tigre* aplicada al ser humano, adquiere rasgos nuevos (*cruel, sanguinario*) derivados del rasgo tomado originariamente del término original (*feroz*).

La figura anterior puede dar una idea aproximada de la complejidad organizativa del significado. Bastaría con pensar en la cantidad

de interconexiones significativas presentes en verbos como *hacer* o *ir*. El hábito de considerar las palabras como unidades independientes de significado nos ha hecho pensar, sin embargo, que la potencialidad comunicativa se circunscribe a la palabra. Pero no es así. Las palabras se relacionan también entre sí, con otras palabras, o con sintagmas y oraciones. Y esa relación mutua se establece a menudo con arbitrariedad, aunque siempre exista alguna causa que la justifica u origina. Un sistema lingüístico así considerado, consta de una verdadera red multidimensional de significados y rasgos semánticos entrelazados cuyo mejor modelo lo encontramos en las conexiones neuronales. De ahí que el modelo de *constelaciones léxicas* propuesto por Cantos y Sánchez (2001) sea el más adecuado para captar la organización de los significados léxicos, tanto en el ámbito de la palabra como en el de las unidades formales más amplias.

La organización de los significados es, pues, compleja e interdependiente. En ello reside su dificultad, pero también las claves que facilitan la comunicación. La diferencia entre los rasgos sémicos de *tigre* (1) respecto a *tigre* (2) y respecto a *animal mamífero* permite a los diccionarios establecer diferencias significativas. A su vez, las interdependencias entre los rasgos que configuran el significado de *tigre* (1) (felino, tipo de piel, garras, ferocidad…) obliga al hablante a seleccionar un determinado tipo de palabras cuando utiliza este significado en el discurso, palabras co-textuales que no son las mismas cuando selecciona el significado de *tigre* (2) en la comunicación. En los ejemplos siguientes pueden apreciarse las diferencias del co-texto:

(8) a. Yo siempre digo que a ver quién le quita las rayas al *tigre*.
b. Distante, se oye el bramido de un toro que tal vez ha venteado al *tigre*.
c. "Un día se la comió el *tigre*", le dijo Estela a Héctor, poniéndole mirada de misterio.
d. Todos los días, cada *tigre* ingiere una cantidad entre seis y ocho kilos de carne de ternera.
e. El culto a San Martín incluye la danza de los cofrades, vestidos con pieles de *tigre* y de venado.

(9) a. Se sospecha que los responsables son los *tigres* tamiles, un grupo guerrillero que lucha por la independencia del norte.
b. Diplomáticos, ministros de Estado, empresarios, militares, altos funcionarios y los habituales *tigres* del cóctel, allí reunidos para celebrar el onomástico del Emperador.
c. Un ejemplo de utilización de estrategias de penetración de mercados, lo constituyen los *tigres* industriales del Sudeste asiático.
d. Con su actual distribución del ingreso, el Perú no podría llegar a ser un *tigre* económico.
e. Golean con el tiro del *tigre* y se apoyan en los palos del poste para tomar impulso a la hora de bloquear el balón.

En términos generales, las palabras del entorno de *tigre* como animal y como hombre son diferentes. El hecho sería mucho más relevante y notorio si en vez de basarnos en cinco frases, tomáramos unas cuantas docenas o cientos para elaborar los listados de frecuencia. Pero véanse, no obstante, las diferencias del entorno léxico de cada significado:

Tabla 5.

Cotexto de tigre (1)	***Cotexto de tigre (2)***
Rayas, bramido, cantidad, carne, cofrades, culto, danza, días, distante, incluye, ingiere, kilos, pieles, ternera, toro, venado, venteado , vestidos, vez, día, comió, cara, mirada, misterio	actual, apoyan, asiático, balón, bloquear, celebrar, cóctel, constituyen, diplomáticos, distribución, económico, ejemplo, emperador, empresarios, estado estrategias, funcionarios, golean, grupo, guerrillero, habituales, hora, impulso, independencia, industriales, ingreso, llegar, lucha, mercados, militares, ministros, palos, penetración, poste, responsables, reunidos sospecha, sudeste, tamiles, tiro, tomar, utilización

No todas esas palabras constituyen la clave necesaria para identificar uno u otro significado, pero en cada oración aparecen las suficientes para llevar a cabo esa tarea. Así, en (8)a. basta con el término *rayas*;

en b. basta con 'bramido del toro, venteado'; en c. es suficiente 'se la comió'; en d. basta con 'ingiere seis u ocho kilos de ternera'; y en e. es suficiente la presencia de 'vestidos, pieles de tigre'.

Y de manera similar, en 9 (a)., el término 'grupos guerrilleros' es suficiente para excluir el significado de 'tigre animal'; en b. lo es 'cóctel, reunidos, empresarios…'; en c. '(tigres) industriales'; en d. '(tigre) económico'; y en e. 'tiro, apoyan, bloquean el balón'.

La razón que explica la dependencia de *tigre* (1) o *tigre* (2) en relación con cada una de las palabras anotadas reside en la necesidad de compartir uno o más rasgos sémicos. Así las 'rayas' son inherentes a la piel del tigre animal; el 'bramido del toro' frente a un tigre real forma también parte del conocimiento que tenemos del mundo circundante; o la 'ingestión de seis u ocho kilos de ternera' sólo puede llevarla a cabo un tigre animal… De otra parte, un 'grupo guerrillero', típicamente constituido por humanos, condiciona el significado de *tigre* a su acepción figurada (aplicada al ser humano); sólo un humano puede 'tirar y bloquear el balón', al igual que sólo los humanos toman un 'cóctel', o son 'empresarios'. En definitiva, la red de significados tejida en torno a las palabras o por las mismas palabras establece una asociación de dependencia entre ellas. Los hablantes disponemos del conocimiento necesario para detectar tales lazos y, en consecuencia, no tenemos ninguna dificultad en desambiguar el significado que en cada caso transmite un término utilizado en el discurso, aunque tal término sea potencialmente polisémico.

Si se admiten las premisas detalladas hasta aquí, será fácil de comprender por qué la DAS debe resolver algunos problemas fundamentales para lograr sus objetivos. Los diccionarios actuales no son instrumentos adecuados para los fines de la DAS o de la TA. Se ha demostrado con suficiente claridad que el significado de las palabras no depende solamente de cada una de ellas, sino que está íntimamente ligado y condicionado por otros elementos del discurso. La dimensión del cotexto y la red de dependencias mutuas están aún por definir. En las muestras investigadas con el fin de aplicar algoritmos basados en el co-texto para discriminar los significados de las palabras en el uso comunicativo se han logrado ya éxitos notablemente esperanzadores (Agirre & Edmonds, 2006). Sánchez, Cantos y Almela (2007), en una muestra realizada sobre 40 términos del inglés y del español han logrado un éxito desambiguador automático de significados de entre el

60% y el 98%; el algoritmo desambiguador se elaboró a partir del poder discriminante del cotexto, habiendo extraído la información necesaria de sendos corpus equivalentes, de 20 millones de palabras cada uno (*Cumbre*, para el español, y *Lacell*, para el inglés).

Lo mencionado respecto al poder desambiguador del co-texto y contexto no debe llevar a la conclusión de que todas las claves para fijar el significado residen ahí. La morfología, la sintaxis y la entonación son también elementos relevantes y a veces decisivos. La DAS, por tanto, debe abordarse desde distintos frentes, entre los cuales el co-texto es uno de los más importantes; de hecho el más poderoso, según lo que sabemos por los estudios que se han llevado a cabo o se llevan a cabo en la actualidad (Agirre y Edmonds, 2006). Además, no debe perderse de vista el hecho de que la definición de qué es exactamente una acepción o cuáles son los significados exactos que tiene cada palabra distan mucho de ser nítidos y precisos (Resnik 2006). En la mayor parte de esos estudios, los corpus desempeñan un papel principal. Una vez más, conviene recordar que la conjunción entre los recursos proporcionados por los corpus y la velocidad de proceso que ya actualmente alcanzan los ordenadores, muestran un camino de investigación prometedora.

La DAS no es un objetivo en sí mismo. Es más bien una tarea intermedia para el logro de otros fines terminales, como son la TA, o la extracción de información más aquilatada de las grandes bases de datos textuales, o el análisis del contenido textual, o incluso el mismo análisis gramatical. La capacidad para identificar los significados correctos de las unidades léxicas en la comunicación es un requisito indispensable para la comprensión del mensaje. Y, por lo tanto, también es una dificultad que la TA ha de superar.

5. Conclusión

Como habrá podido inferirse de lo dicho hasta ahora, la TA, la DAS y los corpus constituyen un trío cuya potencialidad se fundamenta en un denominador común: el tratamiento automático del lenguaje, el cual propicia que las máquinas cobren una 'comprensión' adecuada de los textos que procesan. Este es el principal objetivo que actualmente se persigue en la aplicación de los ordenadores al campo de la lengua.

Apenas han transcurrido 50 años desde que se iniciaron las investigaciones en torno al tema y puede afirmarse que se han logrado algunas metas dignas de mención. En concreto, (i) conocemos mejor los mecanismos que rigen la configuración del significado y (ii) disponemos de recursos que permiten el análisis de grandes bases textuales, capaces de garantizar la fiabilidad de las conclusiones obtenidas, incluidas las de naturaleza semántica. Con este bagaje, ya disponemos de un excelente punto de partida.

En el estadio actual de la investigación, son varios los objetivos concretos que podríamos mencionar como prioritarios. Menciono uno que es de especial relevancia para la TA: la elaboración de una base de datos léxica útil y adecuada para el tratamiento automático del lenguaje. Es algo cuya necesidad se hace cada día más perentoria. Los repertorios léxicos tradicionales (diccionarios) son insuficientes para la nueva época que se abre en el tratamiento del lenguaje por máquinas, y no sólo por humanos. Los nuevos repertorios léxicos serán solo parcialmente semejantes a los actuales diccionarios. El tratamiento automático de la lengua exige, por ejemplo:

i) la ampliación del concepto de unidad léxica como portadora de *unidades de significado*. Las unidades de significado no solamente residen en las palabras en cuanto formas separadas por un espacio en blanco (eso es la palabra para un ordenador y para muchos hablantes), sino también en agrupaciones léxicas en las que pueden estar integradas varias formas lingüísticas, e incluso oraciones enteras. La elaboración de un repertorio de unidades de significado de esta índole requiere de una profunda investigación con corpus lingüísticos, tanto monolingües como bilingües o multilingües.
ii) La creación de un glosario de palabras y unidades de significado en el cual se expliciten no solamente los significados asignados a las formas en la comunicación, sino también los contextos que hacen que se active cada uno de los posibles significados, particularmente en el caso de unidades polisémicas.

La creación de una base de datos léxica de estas características no es un objetivo a corto plazo, sino más bien a medio y largo plazo. Su necesidad para el tratamiento automático del lenguaje no es cuestionable y, por lo tanto, cuanto antes se inicie el trabajo, mejor. De he-

cho ya existen algunas bases de datos léxicos que apuntan en esta dirección. Entre ellas, *Wordnet* es la más conocida. También son conocidas algunas iniciativas privadas útiles, como la de A. Kilgarriff (Kilgarriff et al. 2004), quien ha desarrollado un programa para la elaboración de *perfiles léxicos* (conjunto de datos representativos del cotexto en el que suele usarse una palabra concreta). La base de datos que se precisa debe ser más exhaustiva que las existentes hasta el momento. La máquina necesita disponer de tantos datos como los hablantes cuando utilizan el lenguaje. Deberá ser, además, una base de datos debidamente estructurada, en la que confluyan nuestros conocimientos lexicográficos, morfológicos, sintácticos, semánticos y pragmáticos. Evidentemente, una base de datos de estas características requerirá tiempo y la conjunción de esfuerzos, sin exclusiones, especialmente las que pudieran derivar de la *adscripción* de los investigadores a una determinada teoría lingüística. Deberá ser un trabajo en el que se aúnen y acumulen todos los conocimientos disponibles sobre una lengua, pero organizados de tal manera que sean accesibles al ordenador y puedan ser procesados por él. La TA ganará en eficacia y calidad en la medida en que una base de conocimientos lingüísticos de este tipo pueda estar disponible.

6. Referencias

Abaitua, Joseba. 2002. "Tratamiento de corpora bilingües". En *Tratamiento del lenguaje natural: tecnología de la lengua oral y escrita.* M. A. Martí y J. Llisterri (eds). Barcelona: Fundación Duques de Soria e Edicions de la Universitat de Barcelona. 61-90.

Almela, M. 2006. *From Words to Lexical Units: A Corpus-Driven Account of Collocation and Idiomatic Patterning in English and English Spanish*. Frankfurt am Main / Berlin / Bern / Bruxelles / New York / Oxford / Wien: Peter Lang.

Almela, M., Sánchez, A. y Cantos, P. 2006. "Lexico-Semantic mapping of meanings in English and Spanish: A model of analysis". *Aspects of Translation.* J. M. Bravo (ed). Valladolid: Universidad de Valladolid.11-43.

ALPAC. 1966. *Languages and Machines: Computers in Translation and Linguistics*. A report by the Automatic Language Processing

Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences, National Research Council, 1966. Publication 1416.

Baker M. 1993. "Corpus linguistics and translation studies: implications and applications". En *Text and Technology. In Honour of John Sinclair.* M. Baker, G. Francis y E. Tognini-Bonelli (eds). Amsterdam: Benjamins. 233-250.

Baker M. 1995. "Corpora in translation studies. An overview and some suggestions for future research". *Target* 7 (2): 223-243.

Baker M. 1999. "The role of corpora in investigating the linguistic behaviour of professional translators". *International Journal of Corpus Linguistics* 4: 281-289.

Baker, M., Francis, G. y Tognini-Bonelli, E. (eds). 1993, *Text and technology. In honour of John Sinclair.* Amsterdam: Benjamins.

Bar-Hillel, Y. 1960. "Automatic translation of languages". En http://www.mt-archive.info/Bar-Hillel-1960.pdf

Borin L. (ed). 2002. *Parallel Corpora, Parallel Worlds*. Amsterdam / New York: Rodopi.

Brown, P. F., Lai, J. y Mercer, R. L. 1991. "Aligning sentences in Parallel Corpora". *Proceedings of the Association for Computational Linguistics ACL'91*. Berkeley: 169-176.

Cantos, P. y Sánchez, A. 2001. "Lexical constellations: What collocates fail to tell". *International Journal of Corpus Linguistics* 6 (2): 199-228

Carl, Michael y Way, Andy (eds). 2003. *Recent Advances in Example-Based Machine Translation*. Dordrecht: Kluwer.

Chen, K.H y Chen, H. H. 1995. "Aligning bilingual corpora especially for language pairs from different families". *Informations-Sciences-Applications*, 4 (2): 57-81.

Chan Yee S. y Hwee T. Ng. 2005. "Scaling up word sense disambiguation via parallel texts". *Proceedings of the 205h National Conference on Artificial Intelligence ICAA.* Pittsburgh, USA: 1037-1042.

Chen, S. 1993. "Aligning sentences in bilingual corpora using lexical information". *Proceeding of ACL-93*: 9–16.

Cowie, Joe, Guthrie, A. y Guthrie, L. 1992. "Lexical disambiguation using simulated annealing". *Proceedings of the 14th Inter-*

national Conference on Computational Linguistics COLING-Nantes, France: 359–365.

Gale, W. A. y Church, K. W. 1991. "A program for aligning sentences in bilingual corpora". *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics.* Berkeley, California. 177-184.

Gale, W. A., Church, K. W. Yarowski, D. 1992. "Work on statistical methods for word sense disambiguation". *AAAI Fall Symposium Series: Probabilistic Approaches to Natural Language* Working Notes. Cambridge, MA: 54-60.

Gelbukh, Alexander and Sidorov, Grigori. 2006. "Alignment of paragraphs in bilingual texts using bilingual dictionaries and dynamic programming". *Lecture Notes in Computer Science* 4225. Dordrecht: Springer: 824-833.

Hoey, M. 1991. *Patterns of Lexis in Text*, Oxford: Oxford University Press.

Hoey, M. 2005. *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Hofland K. y Johansson S. 1998. "The Translation Corpus Aligner. A program for automatic alignment of parallel texts". En *Corpora and Cross-Linguistic Research. Theory, Method and Case Studies.* Johansson S. y Oksefjell S. (eds). Amsterdam / Atlanta: Rodopi: 87-100.

Hutchins, J. 1999. "Warren Weaver Memorandum: 50th Anniversary of Machine Translation". *MT News International. Newsletter of the International Association for Machine Translation*. 22 (8-1): 5-6 y 15. http://www.mt-archive.info/MTNI-22.pdf

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

Hutchins, J. 2005. *Milestones in Machine Translation*, No.6: Bar-Hillel and the non feasibility of FAHQT.

Hutchins, W. John y Somers, Harold L. 1992. *An Introduction to Machine Translation*. London: Academic Press.

Kilgarriff A., Rychly, Pavel, Smrz, Pavel y Tugwell, David. 2004. "The Sketch Engine". *Proceedings of the Eleventh EURALEX International Congress.* Lorient, France: 105–116.

Kilgarriff, A. 1993. “Dictionary of word sense distinctions: An enquiry into their nature”. *Computers and the Humanities*, 26: 365-387.

Kilgarriff, A. 2006. “Word Senses”. En *Word Sense Disambiguation. Algorithms and Applications.* Agirre, E., Edmonds, P. (eds). Dordrecht: Springer. 29-45.

Langlais, Ph, Simard, M. y Veronis, J. 1998. “Methods and practical issues in evaluation alignment techniques”. *Proceedings of Coling-ACL-98*.

Laviosa S. 1997. “How comparable can comparable corpora be?”. *Target* 92: 289-319.

Laviosa S. 1988. “The English Comparable Corpus ECC. A resource and a methodology for the empirical study of translation”. En *Translation and Meaning.* Thelen M. y Lewandowska-Tomaszczyk B. (eds). Part 3. Maastricht: Hogeschool Maastricht.

Laviosa S. 1998. “L’approche basée sur le corpus / The corpus-based approach. A new paradigm in translation studies”. *META* 43 (34): 474-479.

Laviosa S. 1998. “The English Comparable Corpus. A resource and a methodology”. En *Unity in Diversity, Current Trends in Translation Studies*. Bowker L., Cronin M., Kenny D. y Pearson, J. (eds). Manchester: St. Jerome Publishing. 101-112.

Laviosa S. 2002. *Corpus-based Translation Studies. Theory, Findings, Applications*. Amsterdam / New York: Rodopi.

Lenssen, Phillip. 2005. “Google Translator: The Universal Language”. *Google Blogoscoped.* http://blog.outer-court.com/archive/2005-05-22-n83.html

Lesk, M. 1986. “Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone”. *Proceedings of the 1986 ACM SIGDOC Conference*. Toronto, Canada. 24–26

McEnery, A. M. y Oakes, M. P. 1996. “Sentence and word alignment in the CRATER project”. En *Using Corpora for Language Research.* J. Thomas y M. Short (eds). London: Longman. 211–231.

McEnery, Tony and Wilson, Andrew. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Nagao, Makoto. 1984. "A framework of a mechanical translation between Japanese and English by analogy principle". En *Artificial and Human Intelligence.* Elithorn A. y Banerji. R. (eds). Elsevier Science Publishers. 173-180.

Nagao, Makoto. 1989. *Machine Translation: How far Can it Go?* Oxford: Oxford University Press.

Olohan, Maeve. 2004. *Introducing Corpora in Translation Studies*, Abingdon: Routledge.

Pierce, John R., Carroll, John B. et al. 1966. *Language and Machines. Computers in Translation and Linguistics*. ALPAC report, National Academy of Sciences, National Research Council, Washington, DC.

Pustejovski, J. y Boguraev, B. 1996. *Lexical Semantics: The Problem of Polysemy.* Oxford: Clarendon.

Resnik, Philip. 2006. "WSD in NLP Applications". En *Word Sense Disambiguation. Algorithms and Applications.* Agirre, E. y Edmonds, P. (eds). Dordrecht: Springer. 299-337.

Sánchez, A., Sarmiento, R., Cantos, P. y Simón, J. (eds). 1995. *CUMBRE. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones*. Madrid: SGEL.

Sánchez A. y P. Cantos. 1997. "Predictability of word forms types, and lemmas in linguistic corpora". *International Journal of Corpus Linguistics* 22: 251-272.

Sánchez, A. y Almela, M. 2006. "Formalización de las correspondencias entre acepciones y contextos sintagmáticos en español e inglés". *Actas del XXXV Simposio Internacional de la Sociedad Española de Lingüística*. León: Universidad de León. 1664-1679.

Sánchez, A., Almela, M. y Cantos, P. 2006. "Lexico-semantic mapping of meanings in English and Spanish: A model of analysis". En *Aspects of Translation.* J. M. Bravo (ed). Valladolid: Universidad de Valladolid. 11-43.

Sánchez, A., Cantos, P. and Almela, M. 2007. "Lexical Constellations and the structure of meaning: A prototype application to WSD". En *Computational Linguistics and Intelligent Text Processing, 8th International Conference, CICLing 2007, Mexico City.* A. Gelbukh (ed). Berlin: Springer Verlag. 275-278.

Sato, S. y Nagao, M. 1990. ”Toward memory-based translation”. *Proceedings of the 13th International Conference on Computational Linguistics.* Coling 90, Helsinki, Finland. 247-252.

Shannon C. E. y Weaver, W. 1949. *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. 2004. *Trust the Text: Language Corpus and Discourse*. London: Routledge.

Sinclair, J. (ed). 1987. *Looking up*. London: Collins.

Stevenson, M. y Wilks, Y. 2001. “The interaction of knowledge sources in word sense disambiguation”. *Computational Linguistics*. 27 (3): 321-349.

Teubert W. 1996. “Comparable or parallel corpora?”. *International Journal of Lexicography* 9: 238-264.

Teubert W. 2002. “The role of parallel corpora in translation and multilingual lexicography”. En *Lexis in Contrast. Corpus-based Approaches.* B. Altenberg y S. Granger (eds). Amsterdam / Philadelphia: John Benjamins. 189-214.

Tufis, D., Ion R. e Ide N. 2004. “Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets”. *Proceedings of the 20th International Conference on Computational Linguistics COLING,* Geneva, Switzerland. 1312-1318.

Véronis, Jean (ed). 2000. *Parallel Text Processing: Alignement and Use of Translation Corpora*. Dordrecht: Kluwer.

Yang, J. y Lange, E. D. 1998. “Systran on AltaVista: a user study on real-time machine translation on the Internet”. En *Machine Translation and the Information Soup*: *Third Conference of the Association for Machine Translation in the Americas, AMTA’98, Langhorne, PA.* D. Farwell, L. Gerber y E. Hovy (eds). Berlin: Springer. 275–285.

Zanettin, F., 1998. “Bilingual comparable corpora and the training of translators”. *Meta*, 43 (4): 616-630.