

El sistema de traducción asistida por ordenador MicroCAT. Sus ventajas y limitaciones

J. Gabriel de Amores Carredano
Universidad de Sevilla

1. Descripción del Sistema

Como su nombre indica, MicroCAT es un sistema de traducción asistida por ordenador («Computer Assisted Translation») desarrollado por la Compañía Weidner en 1978. Sus creadores fueron un grupo de mormones a los que fascinaba la idea de traducir sus documentos religiosos al mayor número de lenguas posibles.

Actualmente se dispone de los siguientes pares de lenguas: inglés-español, español-inglés, inglés-francés, inglés-alemán, inglés-portugués e inglés-italiano. Se están desarrollando además las versiones francés-inglés, alemán-inglés, alemán-francés, francés-español y japonés-inglés.

El programa corre por el momento en los modelos XT y AT de IBM y en los PC de ITT. No obstante, se está estudiando la posibilidad de ampliar la oferta de equipos que pudieran admitirlo. La velocidad de traducción es aproximadamente de 1.600 palabras por hora.

Los textos se pueden introducir manualmente en el procesador de textos del sistema, o por medio de un lector óptico («scanner») que produce el texto en formato ASCII, Wordstar o WordPerfect. Mediante un programa de conversión se puede cambiar el formato de los textos al del procesador de textos de MicroCAT.

El menú principal consta de 8 opciones

a) Procesador de textos: usado para introducir el texto que se va a traducir y para pulir la versión en bruto que genera el sistema. Es muy parecido a WordPerfect, pero ofrece algunas ventajas adicionales, como edición de las dos versiones (Texto Fuente y Texto Destino) sincrónicamente con la pantalla dividida.

b) Búsqueda de palabras: crea un fichero con las palabras que no se han

encontrado en el texto.

c) Actualización del diccionario: es lo más importante del sistema. Permite al usuario introducir entradas simples o compuestas en uno o varios diccionarios, mediante la codificación de las características gramaticales de los elementos del discurso.

d) Utilidades de diccionario: permite crear diccionarios de términos específicos, y organizarlos secuencialmente según una jerarquía.

e) Traducción: traduce el texto fuente automáticamente, produciendo un texto destino que habrá de ser pulido por el traductor humano.

f) Traducción inmediata: se usa para comprobar la validez de las entradas creadas en el diccionario, admitiendo una línea completa como máximo.

g) Procesamiento retardado: permite definir una serie de trabajos (traducción o búsqueda de palabras) que el programa realizará cuando se lo ordenemos (p.ej. por la noche) para ser post-editados a la mañana siguiente. Admite hasta 14 textos distintos.

h) Contador de palabras: el objeto principal es contar las palabras traducidas por texto.

2. *Objetivos del sistema*

Como explicaremos a continuación, el programa no pretende producir un texto en alta calidad. Está más bien orientado a gabinetes de traducción de textos técnicos. En la mayoría de los casos el cliente sólo exige consistencia terminológica, rapidez en disponer del producto, y bajo costo. Lo que obviamente el sistema no puede generar es estilo literario.

Por otra parte, existe una gran dependencia entre la calidad del texto fuente y la del texto destino. Es algo sabido que, no pocas veces, los textos de materias técnicas suelen ofrecer menor claridad estilística, e incluso errores en el uso de las estructuras gramaticales.

Se deduce, pues, que hace falta algún tipo de preedición, en la medida en que vamos conociendo mejor el programa y prevemos que ciertas secuencias van a producir confusión. Son particularmente útiles la incorporación de artículos delante de los sustantivos para ser distinguidos de sus homógrafos verbales, la inclusión de comas en caso de acumulación de varios adjetivos delante del sustantivo y cierta ayuda para delimitar el ámbito de la negación y resolver adecuadamente algunos tipos de coordinación.

P. ej. en el *texto fuente* que proponemos como ejemplo, se han incluido comas delante de «hard disk»: «single, large, hard disk» y «multiple, smaller, hard disks».

De todo esto podemos concluir que el sistema ha sido diseñado con vistas a su utilización dentro de un marco comercial, para textos técnicos, en los

que se puede distinguir un sublenguaje que facilita la reducción de la polisemia. A diferencia de la traducción literaria, la traducción técnica persigue la fidelidad al original mediante el trasvase de información como principal objetivo.

En no pocas ocasiones, el lector especializado en un área técnica queda plenamente satisfecho con traducciones producidas por sistemas de TA. Él podrá suplir adecuadamente con su conocimiento sobre el tema, lo que falte al texto en cuanto a calidad se refiere.

El traductor humano especializado es caro, lento y su dominio de la terminología actual es limitado. Vemos, pues, que la traducción por ordenador (ya sea traducción automática o traducción asistida por ordenador), va a jugar un papel primordial en el campo de la traducción técnica.

3. El Proceso de Traducción

El sistema opera según los siguientes pasos:

El manipulador de entradas («Input Handler»), controla el texto en Lengua Fuente (LF) y lo divide en palabras y oraciones; la consulta de diccionario («Dictionary Lookup») lleva a cabo el análisis morfológico y examina el diccionario: durante este proceso, el sistema busca las palabras no conocidas y crea un fichero con las unidades léxicas independientes no encontradas. Posteriormente, el transformador de expresiones idiomáticas («Idiom Converter») identifica las locuciones idiomáticas y compuestos nominales de hasta ocho unidades léxicas. A continuación tienen lugar las fases de resolución de homógrafos («Homograph Resolution»), por medio de la consulta de contextos incluidos en el diccionario y el análisis sintáctico («Syntactic Analysis»), para la identificación de sintagmas nominales y verbales, así como las posibles interrelaciones, para producir cierto tipo de análisis de estructura sintagmática.

Las fases finales de transferencia y síntesis incluyen rutinas para la reordenación del texto de acuerdo con las estructuras de la LD, conjugación verbal y concordancia, rutinas de retoque para insertar los artículos en LD y las preposiciones necesarias, según la información del diccionario.

Por último, el manipulador de salida («Output Handler») reorganiza el texto de acuerdo con el formato original LF. A continuación, el texto pasa al corrector de texto («Text Amender»), un procesador de textos para edición.

4. Relación usuario-sistema

En todo tipo de programas para ordenadores se procura que el usuario y

la máquina mantengan una relación lo más cordial posible. Para ello es fundamental un conocimiento mutuo que debe asentarse también sobre el conocimiento de las propias limitaciones. No podemos pedir al sistema que haga algo para lo cual no ha sido diseñado, o para lo que no posee la información adecuada.

En este sentido, el usuario llega a tener la impresión de depender del sistema en gran medida. Por una parte, el usuario habitual debe saber exactamente cómo están definidas las entradas, qué términos posee ya, cuáles son locuciones idiomáticas y de qué tipo. En segundo lugar, el usuario no tiene acceso a las reglas gramaticales. Anandiou comenta al respecto¹:

«Without any knowledge of the linguistic rules employed, and being constrained to use a menu-driven dictionary up-date module, the user can only make guesses at the eventual behaviour of the system when he gives some 'most-like' response».

Podemos añadir que el usuario puede llegar a tener la sensación de estar trabajando siempre de cara al futuro. Es decir, el trabajo de incluir nuevos términos y entradas en los diccionarios tendrá su fruto en posteriores textos en los que aparezcan esas mismas unidades léxicas, para ir reduciendo progresivamente el trabajo de actualización de los diccionarios de un sublenguaje específico.

5. La gestión de diccionarios

Constituye el eje del programa. En este tipo de sistemas, la información de los diccionarios es de capital importancia. Cuanta más información contenga una entrada, mayor probabilidad de éxito tendremos.

El diccionario base («Core») varía en extensión según el par de lenguas de que se trate. La media es de 13.000 entradas. Este diccionario no se puede modificar. Para cambiar algo, hay que copiar la entrada en otro diccionario y hacer allí la modificación deseada.

Como ya hemos dicho, el usuario puede crear cuantos diccionarios crea oportuno, y organizarlos a su arbitrio. La ventaja de esta jerarquía está en que la búsqueda de términos va de lo más particular a lo más general.

¹ Anandiou, S. «A Brief Survey of Some Current Operational Systems» en Margaret King, (ed) (1987), *Machine Translation Today*. Edinburgo: Edinburgh University Press, pp. 171-192.

- 1^º diccionario(s) específico(s)
- 2^º diccionario general
- 3^º diccionario base

El usuario puede incluir entradas en los diccionarios en su forma canónica, o como locuciones idiomáticas de hasta 8 unidades. Además existe la posibilidad de incluir locuciones de concurrencia gramatical posible (*may match*) u obligatoria (*must match*), es decir, formas que pueden o deben ir seguidas siempre de determinado elemento del discurso. P.ej. 'ALL OF %', donde %, siempre que sea un pronombre personal plural (*us, you, them*) --> «todos nosotros, vosotros, ellos», o 'VERY %' donde % siempre que sea sustantivo --> «mismo %», en vez de «muy».

Durante el proceso de actualización de los diccionarios hay que «olvidarse» un poco de los criterios estrictamente lingüísticos en determinadas ocasiones. Sin acceso a las reglas lingüísticas usadas, difícilmente podemos apostar por que una entrada funcionará correctamente en el 100% de los casos. Es el momento de ensayar diversas posibilidades -aunque se cometan desviaciones con respecto a la lingüística teórica- hasta dar con aquella que solucione aceptablemente el mayor número de casos. La estadística es algo muy importante en este tipo de programas. Debido a la limitación de posibilidades que se admiten como traducción de cada entrada, habremos de considerar previamente si cierta entrada con cierta traducción nos resultará o no rentable en el futuro.

6. Actualización de los diccionarios

MicroCAT tiene definidos 10 elementos del discurso: verbo, sustantivo, adjetivo, pronombre, preposición, conjunción, interjección, cuantificador y determinante.

El usuario participa interactivamente con el sistema para introducir nuevas entradas, y caracterizarlas por medio de códigos (flags), que responden al paradigma requerido: flexión nominal y adjetival de género y número, si el adjetivo y el adverbio han de situarse en la traducción en la misma posición con respecto a su término: p.ej. «some», a diferencia de otros adjetivos, como «green», debe permanecer en la traducción en el mismo lugar (*stay put*).

some cars ---> algunos coches
green cars ---> coches verdes

o si están en forma comparativa o superlativa (*better, most, faster, fastest...*)

Sólo el sustantivo contiene cierta información semántica. El menú es: humano, grupo, parte del cuerpo, animado (no humano), inanimado, contable e incontable, si es un nombre propio o un sustantivo que indica lugar o tiempo.

Las distintas caracterizaciones influirán en la inserción o no de artículo:

I gave my brother a book --> Yo le di *a* mi hermano un
[+Human] libro
Committee decisions --> las decisiones *del* comité
[+Group]
Laughing cow --> la vaca *que* rie
[+Animate]
A statue of brass --> una estatua de \emptyset bronce
[+Mass]
History of Law --> historia *del* derecho
[+Abstract]

El adjetivo presenta asimismo varias posibilidades:

* El adjetivo rige subjuntivo:

It is important that they arrive tomorrow --> Es importante que lleguen mañana.

It is not probable that he will attend the meeting --> No es probable que asista a la reunión.

* Adjetivo «estar»:

He is tall --> Él es alto.

He is ill --> Él está enfermo.

El verbo contiene información sobre conjugación (88 paradigmas distintos para el español), valencia, transitividad, reflexividad, si en la negación rige un subjuntivo (p. ej. «no creo que venga hoy»), tipo de objeto directo e indirecto, y si el participio presente («-ing») y pasado («-ed») pueden ser adjetivos.

El elemento más complejo es el verbo. Todo verbo puede generar la forma de adjetivo «-ed» y la forma «-ing» a no ser que se restrinjan las posibilidades. Por definición, todos los verbos son tratados como homógrafos.

Uno de los problemas principales que se plantean en la TA es el tratamiento de homógrafos. MicroCAT incorpora una rutina de tratamiento de homógrafos tras la consulta de palabras y tratamiento de locuciones idiomáticas. Existen 12 tipos de homógrafos:

verbo solo: genera las formas adjetivas «-ed» e «-ing»:
achieve.

verbo y sustantivo: bank
sustantivo y adjetivo: concrete
verbo, sustantivo y adjetivo: level
verbo, sustantivo y adverbio: fast
sustantivo, adjetivo y adverbio: right
verbo, sustantivo, adjetivo y adverbio: back
sustantivo y adverbio: upstairs
adjetivo y adverbio: only
verbo y sustantivo en «-ing»: building
adjetivo y adverbio: only
verbo y adjetivo (no «-ed» ni «-ing»): direct
verbo y adverbio: please

En esta fase se dice que el sistema compara la forma propuesta y decide, según el contexto de la oración, a qué elemento del discurso pertenece entre los posibles. El problema es que falla en un porcentaje alto, ante el asombro del usuario. Sin lugar a dudas, es la parte más compleja del programa. Hasta que se llega a dominar perfectamente pueden pasar varias semanas, ya que el número de combinaciones posibles es bastante alto.

7. Un ejemplo de traducción inglés-español

Exponemos a continuación un texto breve extraído de un manual de ordenadores (ITT XRT). En la versión 1 se puede ver el resultado con la consulta del diccionario base solamente. La versión 2 ofrece el texto de salida tras la modificación de los términos que nos han parecido relevantes, e incluirlos en un diccionario específico llamado «COMPUTERS». La versión 3 es la resultante de haber hecho lo mínimo posible de post-edición. El resultado es aceptable.

TEXTO FUENTE

Your computer supports a number of different hard disks, floppy disks, and streaming-tape drive combinations (Figure 3-1). Additionally, you can divide or partition your hard disk into sections that make your single, large, hard disk appear as multiple, smaller, hard disks. For example, you might divide your hard disk into two partitions, one reserved for XENIX and the other reserved for MS-DOS. Or you might dedicate the entire disk to XENIX.

VERSION 1

Su ordenador admite los discos duros, diferentes y varios, discos flexible, y las combinaciones de unidad de cintas de correr (la Figura 3-1).

Adicionalmente, usted puede dividir o dividir su disco duro en las secciones que hacen su solo disco duro grande aparece los discos duros, más pequeños y como múltiples. Por ejemplo, usted quizás divida su disco duro en dos divisiones, un reservado para el XENIX y el otro reservado para el MS-DOS. O quizás dedique el disco entero al XENIX.

VERSION 2

Su ordenador admite los discos duros, diferentes y varios, discos flexibles, y las combinaciones de la unidad de cinta (la Figura 3-1). Adicionalmente, usted puede dividir o dividir su disco duro en las secciones que hacen su solo disco duro grande aparece los discos duros, más pequeños y como múltiples. Por ejemplo, usted quizás divida su disco duro en dos divisiones, un reservado para el XENIX y el otro reservado para el MS-DOS. O usted quizás dedique el disco entero al XENIX.

VERSION 3

Su ordenador admite diferentes combinaciones de discos duros, flexibles y de unidad de cinta (Figura 3-1). Adicionalmente, usted puede dividir su disco duro en secciones que hacen que en el único disco duro aparezcan varios discos duros, más pequeños y múltiples. Por ejemplo, usted puede dividir el disco duro en dos, uno reservado para el XENIX y el otro reservado para el MS-DOS. O puede dedicar el disco entero al XENIX.

8. Conclusiones

De lo visto podemos sacar conclusiones para todos los gustos, dependiendo de las necesidades y exigencias del usuario.

Siendo objetivos, habría que decir ue el programa es bastante potente si se tiene en cuenta que corre en un Ordenador Personal². Con una actualización inteligente de los diccionarios se puede conseguir una alta rentabilidad, comercialmente hablando.

Los errores más usuales se dan en la concordancia, en la coordinación, en el tratamiento de los modales y en los homógrafos, especialmente aquellos que

² Otros programas comerciales, como p. ej. LOGOS, necesitan ordenadores de tipo medio («mainframe»), con una capacidad mínima de memoria de 4 megas para cada par de lenguas.

pertenecen a campos semánticos distintos: I found the Club (¿pasado de *find* o presente de *found*?)³.

El problema fundamental que plantea este programa está en la dificultad de ser mejorado. Al ser un sistema directo de TA⁴ de la segunda generación, existe poca separación entre los algoritmos, las gramáticas y las bases de datos (diccionarios), con lo que un cambio en cualquiera de estos módulos produce una alteración en la arquitectura de todo el sistema.

Por otra parte se notan diferencias de estilo en algunas traducciones, dada la orientación latinoamericana del programa, o al menos, de los investigadores que lo desarrollaron.

I saw him --> Yo lo vi

9. Posibles aplicaciones del sistema

En nuestro caso concreto -un Departamento de Lengua Inglesa- pensamos que el sistema presenta grandes limitaciones pero, a la vez, grandes posibilidades en dos campos concretos.

Por una parte, nos sirve como apoyo a una asignatura optativa de traducción. La TA es bastante desconocida en nuestro país hoy en día por la mayoría de los lingüistas. Pensamos que una primera aproximación teórico-práctica a la TA, sus posibilidades y limitaciones, abre nuevas puertas a los futuros licenciados en filología inglesa que se quieran dedicar a la traducción. Cada vez son más las compañías y gabinetes de traducción que utilizan este tipo de ayudas a la traducción (diccionarios electrónicos en línea, procesadores de texto, diccionarios de sinónimos...). MicroCAT agrupa todas esas posibilidades, con la ventaja adicional de que se pueden separar convenientemente los distintos campos técnicos que se vayan a traducir con la oportuna gestión de los diccionarios específicos.

³ Se da el caso de que para generar el pasado de *find*, que estadísticamente hablando es más productivo que *found*, habría que escribir **finded*. Lo mismo ocurre con la forma *left*; si queremos generar el pasado de *leave*, habremos de escribir **leaved*, ya que lo normal es que genere «izquierda». La dificultad que entraña la resolución de este tipo de homógrafos es quizá uno de los retos más serios que tiene planteada la TA, y se ha convertido a su vez en una de las razones de mayor peso para introducir técnicas de Inteligencia Artificial en los algoritmos de resolución de homógrafos.

⁴ Para una comprensión acerca de los distintos sistemas de TA, ver Hutchins (1986: 174-78) y Niremburg, ed. (1987: 22-28).

En segundo lugar, sirve también como iniciación a la lingüística computacional. El sistema es fascinante en este sentido, ya que hemos de cambiar nuestro modo de pensar, tratar de «trucar» el programa para que haga lo que realmente queremos, dentro de sus posibilidades. Facilita asimismo una aproximación al diseño de diccionarios electrónicos y a la manera en que se llevan a cabo los procesos de análisis, transferencia y generación de un sistema de TA.

BIBLIOGRAFIA

Grishman, R. (1986). *Computational Linguistics*. Cambridge: Cambridge University Press.

Harris, M.D. (1985). *Introduction to Natural Language Processing*. Londres: Prentice Hall.

Hutchins, W.J. (1986). *Machine Translation: Past, Present, Future*. Chichester: Ellis Horwood Ltd.

King, M. (ed.) (1987). *Machine Translation Today: the State of the Art*. Edinburgo: Edinburgh University Press.

LOGOS: Translation Systems. Logos Computer Systems Deutschland GMBH. Frankfurt. Documentación técnica facilitada al autor de este artículo.

Niremburg, S. (ed.) (1987). *Machine Translation: Theoretical and Methodological Issues*. Cambridge: Cambridge University Press.

Slocum, J. (ed.) (1988). *Machine Translation Systems*. Cambridge: Cambridge University Press.

Weidner Communications Corporation: Computer-Aided Language Translation System MicroCAT. Weidner Translation (Europe) Ltd. Southampton. User's Manual.