

Fuzzy matches, friends or foes? (¿Amigos o enemigos?)

José María Montero Vázquez*

Resumen: El concepto de *fuzzy match* o concordancia parcial ha supuesto una ayuda inestimable en el entorno de la traducción técnica con las herramientas TAO, pero, a su vez, este concepto ha sido aprovechado por las agencias de traducción para imponer criterios de facturación basados en estas concordancias. Su utilidad puede resultar evidente para cualquier traductor técnico, pero a veces esconde verdaderas trampas contra el traductor poco avisado.

Palabras clave: concordancias parciales, traducción asistida por ordenador, TAO, memoria de traducción.

Fuzzy matches, friends or foes?

Abstract: The concept of fuzzy matches emerged as a priceless help for technical translators working with CAT applications, but at the same time is used by the translation agencies to impose their invoicing criteria, based on these type of matches. Their usefulness becomes obvious for any technical translator, however it can be sometimes tricky for the little brisk translator.

Key words: fuzzy matches, computer-assisted translation, CAT, translation memory.

Panace@ 2009; 10 (29): 24-26

Mi primer contacto profesional con los *fuzzy matches* o concordancias parciales se remonta a la última década del siglo pasado (sí, está bien, no queda tan lejos, pero la afirmación no deja de ser cierta). Probaba por aquel entonces Translator's Workbench en su versión 1.0 y no dejaba de sorprenderme la simplicidad del concepto y las ventajas de esa herramienta para mi trabajo. Desde entonces, la filosofía básica operativa de los programas, la traducción asistida por ordenador o TAO,¹ no ha cambiado mucho, pero sí el entorno, las capacidades auxiliares y los complementos aparecidos alrededor de dichos programas. Incluso el concepto de traducción automática se basa en principios lingüístico-computacionales similares a los empleados por estos programas de TAO, aunque los programas de TA (traducción automática) llevan estos principios y características de trabajo a niveles más complejos e introducen, además, diversos tipos de análisis en su programación: análisis sintácticos, semánticos e incluso morfológicos.

Lamentablemente, la lógica computacional subyacente que posibilita la comparación sintáctico-gramatical de dos segmentos en la lengua origen entre sí y su relación con las traducciones encontradas en la base de datos en lengua(s) destino que se emplean en estos programas me es totalmente desconocida, y no pretendo siquiera esbozar en este breve análisis sus conceptos básicos, sino más bien tratar una parte del resultado visible y «aprovechable» de la aplicación de dicha lógica: las coincidencias parciales (*fuzzy matches*). Pero para dar mi opinión sobre ellas deberíamos saber en qué consisten y cuál es su funcionalidad, o al menos la que los creadores, programadores y lingüistas computacionales que idearon y desarrollaron su concepto les dieron en un principio.

Friends?

En su origen, los programas de traducción asistida por ordenador comenzaron siendo almacenes de información en una base de datos sencilla dotada de un motor de búsqueda, es decir, el concepto de memoria de traducción, cuya estrategia de uso principal estriba en el acceso a dicha base de datos para reutilizar y/o adaptar los elementos almacenados.

Los elementos lingüísticos que van a formar la memoria de traducción son, generalmente, oraciones completas, a veces estructuras menores (sintagmas o palabras) y otras veces mayores (párrafos). Dichos elementos se organizan en una base de datos donde se guardan por parejas² de idiomas que siempre están vinculadas. Cuando el motor de búsqueda recupera uno de los constituyentes, que a partir de ahora denominaremos *segmentos*, la parte vinculada se muestra, para que el usuario la reutilice tal cual (*coincidencia total*) o para que la modifique y adapte al nuevo elemento de traducción (*coincidencia parcial*).

Como ya he señalado, los entresijos de la lógica computacional y lingüística empleados en este proceso son complejos y están basados en reglas de la lingüística computacional, pero el resultado es un aprovechamiento parcial de las estructuras traducidas con anterioridad, que pueden modificarse para adaptarlas a la nueva estructura que se traduce, tras lo cual se almacenarán en la memoria de traducción (la base de datos) como una nueva unidad (segmento) y podrán, a su vez, recuperarse y aprovecharse en el futuro. Este modelo de trabajo aumenta exponencialmente el número de posibles reutilizaciones, siempre y cuando se cumplan una serie de condiciones en los textos con los que se trabaja.

Por supuesto no todos los textos tienen «aptitudes» para aprovechar estas coincidencias parciales, y mucho menos las

* Traductor técnico y médico, Sevilla (España). Dirección para correspondencia: jmv@arrakis.es.

coincidencias totales. Para sacarles ventaja, el texto debería ser mínimamente repetitivo, característica muy frecuente en ciertos tipos de documentación técnica que contienen estructuras gramaticales simples y poco dadas al lenguaje florido. También es cierto que con el tiempo he llegado a apreciar otras funciones de las memorias de traducción, como la «concordancia», que convierten este tipo de programas en algo muy útil para cualquier traducción en la que se requiera una mínima coherencia en el vocabulario técnico.

El hecho de que la mayoría de los programas actuales permitan configurar el nivel de coincidencia parcial recuperable por el motor de búsqueda únicamente nos garantiza, si es la opción elegida por el traductor, recuperar segmentos cuya coincidencia sea muy elevada (o absoluta), de tal forma que los cambios necesarios sean mínimos; todos los demás segmentos deben traducirse ab initio. Este planteamiento puede ser adecuado para textos poco repetitivos y que no necesariamente presenten un lenguaje simplista o poco elaborado, en el que el traductor puede traducir con mayor libertad frases y párrafos sin sentir las «ataduras» de los aprovechamientos.

Considero que este sería el uso beneficioso de las coincidencias parciales. Pero seamos sinceros: en un entorno de trabajo en el que las agencias de traducción han decidido imponer criterios de reducción de precios de forma inversamente proporcional al índice de coincidencia, esta forma de trabajo se antoja poco habitual en textos técnicos, a no ser que el traductor trabaje de forma directa con el cliente.

Foes!

Lo habitual acaba siendo que la agencia de traducción analice el texto previamente y compare las unidades de traducción del documento con los segmentos almacenados en su memoria de traducción. Dicho análisis refleja, tanto porcentualmente como en valores absolutos, el nivel de coincidencia total y parcial de dicho documento con las traducciones previas de ese documento u otros similares.³

Como suele ocurrir con todo en esta vida, siempre que algo, ya sea una moda, un programa o una bebida, se imponen al resto de sus competidores se crea una especie de monopolio, y la «tiranía de lo habitual» termina por perdurar. ¿Quién pide un «refresco de cola»? ¿Por qué se compran muchos ordenadores con el sistema operativo Windows? Como este aspecto de la vida no podía ser diferente, Trados® Workbench, aun cuando no fue la primera, ni posiblemente la mejor de las herramientas TAO, ha terminado por imponerse en el mercado que componen las agencias de traducción y los traductores *freelancer* (autónomos), y con ella su cuantificación (y la lógica subyacente) de las coincidencias.⁴

Translator's Workbench analiza el texto que ha de traducirse e informa cuantitativamente del número de palabras y segmentos que tienen una coincidencia absoluta o parcial con los segmentos de la memoria de traducción. La coincidencia absoluta se cuantifica como coincidencia al 100 %, mientras que una coincidencia parcial será clasificada en varios grupos, dependiendo del porcentaje de similitud encontrado con respecto del segmento comparado: del 95 al 99 %, del 85 al 94 %, del 75 al 84 %, del 50 al 74 %, y finalmente, todos

aquellos segmentos cuyo análisis revele una similitud inferior al 50 % serán clasificados como *No match* (sin coincidencia).

Esta estrategia de cuantificación ha sido la excusa perfecta para que las agencias de traducción impongan sobre los traductores autónomos⁵ una reducción de la tarifa inversamente proporcional al valor porcentual de coincidencia, o lo que es lo mismo: a mayor coincidencia, menor es la tarifa aplicable a las palabras que componen dicho segmento.

En ese momento crucial del proceso de gestión de la traducción, una coincidencia parcial con un valor muy bajo⁶ de coincidencia puede convertirse en un enemigo del traductor. Incluso las que presentan un valor alto pueden ser verdaderos traidores a la causa de una traducción adecuada y precisa. Desgraciadamente, no tengo conocimientos de otros idiomas como para extraer conclusiones similares, pero en la pareja de trabajo inglés-español existen claramente infinidad de ejemplos en los que el nivel de coincidencia entre segmentos resulta irrelevante, e incluso engañoso con respecto al tiempo y trabajo empleados en la modificación y retraducción de dicho segmento. En estos casos, las coincidencias parciales se convierten en enemigos declarados del traductor, porque debe pasar un tiempo importante, que no se compensa económicamente, modificando la traducción para adecuarla al nuevo contexto lingüístico.

Mi larga experiencia me ha llevado a aplicar, siempre que las agencias me lo han permitido (es decir, cuando en la agencia se han parado a pensar un poco y a reflexionar sobre lo propuesto y no simplemente a rechazar la propuesta), una solución salomónica: yo establezco tres niveles (como máximo) para mi tarifa: la tarifa base, que se aplica a las palabras nuevas (*No matches*), el 60 % de mi tarifa base, aplicado a TODAS las coincidencias parciales (desde el 50 al 99%), y, por último, el 25 % de la tarifa base, aplicado a las coincidencias totales y las repeticiones internas del texto. Posiblemente, si se hicieran cálculos estadísticos, podría razonarse una postura frente a la otra; sin embargo, mi experiencia me dice que este sistema suele mostrarse bastante «neutro» en cuanto a quién «sale ganando»: es cuestión de plantear una solución simple a una filosofía de trabajo compleja, que posiblemente nunca contente a todas las partes.

Todas estas consideraciones se basan en la creencia en la «bondad intrínseca» de las traducciones pasadas (o en la aceptación de este supuesto), es decir, que se supone que la traducción que podemos aprovechar desde la memoria es precisa y correcta. Si este factor de bondad desaparece, las coincidencias parciales pasan de ser un posible enemigo a convertirse en una pesadilla, y podría darse el caso —no sería la primera vez— de que fuera mejor traducir desde cero que reutilizar el segmento que «amablemente» nos ofrece el programa, aun a riesgo de que el cliente decida no pagarnos esas palabras como debiera.

Está claro que, con la globalización del trabajo, las agencias utilizan una estrategia de «diversificación» de los recursos en lo referente a sus traductores; o sea, cuentan con más de un traductor por cada memoria de traducción e idioma, lo que nos lleva al «uso compartido» de las memorias. El problema surge cuando la agencia envía a un traductor una me-

moria de traducción en la que existen segmentos traducidos y modificados por personas distintas del propio traductor y este ha de enfrentarse al dilema de revisar o aceptar todos los segmentos, aun cuando dichos segmentos puedan incorporar incorrecciones o imprecisiones, debido a la política de tarificación y trabajo que impera en la actualidad. No son muchos los traductores que pueden convencer a las agencias para que paguen una revisión completa de la memoria (que en muchos casos cuenta con miles de segmentos) o que accedan a revisar las tarifas en función de la «calidad contrastada» de la memoria. En este escenario suelen ser la profesionalidad del traductor y el grado de confianza de la agencia en sus cualidades los que imperen, pero también suele darse la situación contraria, en la que el traductor debe dejar pasar ciertas traducciones por falta de tiempo y falta de «colaboración» de la agencia. Será la experiencia personal del traductor la que le indique qué debe hacer en cada situación particular.

Caso aparte son aquellas memorias en las que únicamente ha participado el propio traductor, bien por ser el único traductor asignado, bien debido a que la creación, gestión y ampliación de la memoria corren por su cuenta: clientes directos o traductor único asignado al proyecto, o bien memoria de creación ex profeso para el proyecto en curso. En estos casos, las coincidencias parciales suelen ser de mucha más ayuda, ya que el traductor confía mucho más en ellas que cuando desconoce su origen o ha comprobado que la calidad de las traducciones contenidas en la memoria no es buena. Este escenario debe redundar en beneficio del traductor, no solo en sentido económico, sino también, y tan importante como lo anterior, en la calidad de la propia traducción, ya que estamos seguros de que lo reutilizado tiene la misma calidad que las traducciones recién generadas. El aspecto económico se explica por dos motivos: primero, una mayor velocidad de traducción al reutilizar segmentos anteriores; y segundo, una mayor velocidad de revisión, al detenerse menos tiempo en los segmentos con coincidencia elevada.

False friend foes

Como conclusión, y a modo de resumen, debo decir que, si bien existen casos en los que las coincidencias parciales pueden suponer un riesgo para el traductor poco avisado o falto de experiencia debido a la posible introducción de errores gramaticales en la traducción, la norma es que dichas coincidencias, una vez bien configurada la lógica que las recupera y clasifica, solucionen en parte algunos problemas de

traducción y hagan que esta sea más rápida y precisa, y con ello aumente la productividad del traductor técnico.

Notas

1. *Traducción asistida por ordenador (TAO)* es el término por el que se conoce en España la traducción llevada a cabo con la ayuda de programas informáticos, ya sean las denominadas memorias de traducción o los editores de recursos de *software*.
2. La relación de pareja no siempre es unívoca (1:1), es decir, 1 segmento en idioma origen y 1 segmento en idioma destino, sino que es posible configurar y utilizar relaciones polifuncionales en cuanto al idioma destino, de manera que se puede trabajar con un idioma origen y varios idiomas destino simultáneamente. También se puede introducir una especie de «polisemia» del segmento de lengua origen: un mismo segmento del original puede almacenarse con varias posibles traducciones, y el traductor debe decidir cuál de ellas es la correcta en cada situación.
3. Se introduce aquí un nuevo punto de debate que, aunque se aleja del hilo conductor de este artículo, no deja de ser interesante para muchos traductores técnicos: ¿hasta qué punto es lícito cobrar/pagar menos por las coincidencias parciales? ¿Hasta qué nivel debería el traductor reivindicar que le paguen a la tarifa normal las coincidencias parciales?
4. Mi «relación laboral» con otras herramientas de TAO, como son *Déjà Vu*, *Start Transit*, *Wordfast*, etc, es escasa en comparación con la que mantengo con *Translator's Workbench*, y aunque existen diferencias en el tratamiento de las coincidencias parciales entre estos programas, todos ellos mantienen la filosofía esencial de trabajo de la reutilización voluntaria de las traducciones. Es posible que un análisis comparativo entre estas herramientas revele pequeñas diferencias en los valores absolutos de coincidencias para un mismo texto, pero ello puede deberse al tratamiento informático que cada una de estas herramientas hace de ciertas partes de la oración (determinantes, numerales, etc).
5. Habría que saber si esa misma reducción se aplica a la tarifa que dichas agencias cobran a sus clientes, o al menos si se impone una reducción tan drástica e «irreversible» como la que se nos impone a los traductores.
6. Cabría recordar que los principales programas de traducción asignan valores porcentuales al nivel de coincidencia entre el segmento que se ha de traducir y los segmentos almacenados: un valor elevado de la coincidencia parcial significa una similitud alta entre ambos segmentos, y cuanto más alto sea ese nivel, más bajo será el precio ofertado para las palabras que componen dicho segmento.

