Memorias de traducción en TMX compartidas por Internet

Joseba Abaitua, Universidad de Deusto

Introducción

En traducción automática es un lugar común afirmar que la calidad está reñida con la cobertura, es decir, que es más o menos viable desarrollar sistemas que traduzcan en ámbitos restringidos (como los partes meteorológicos), pero muy complicado, si no imposible, ampliar el ámbito sin perder calidad. La mayoría de los analistas sostienen que este problema es casi insalvable, dada la variedad de estilos, registros, interpretaciones, etc. que pueden darse en los textos sin restricciones. El mercado actual de software de traducción refleja además esta situación perfectamente. Sin embargo, la comunidad científica dispone de los conocimientos teóricos y tecnológicos para que esta limitación deje de serlo pronto. El problema que queda por resolver no es de índole científica ni tecnológica, sino logística. La solución fue sugerida recientemente en *Language International* 10.6 por Minako O'Hagan, autora de *The coming industry of teletranslation*, y consiste en convertir Internet en una inmensa memoria de traducción.

En este artículo voy a hablar de las condiciones que deberían darse para que esto fuera posible. Para ello, en primer lugar voy a analizar el concepto de equivalencia en traducción. Voy a contradecir algunos de los supuestos más extendidos entre los especialistas para proponer una visión alternativa más amplia. A continuación se presentarán algunas de las nociones básicas de la tecnología de memorias de traducción en TMX y aportaré algunos ejemplos. Finalizaré repasando conceptos surgidos en el campo del desarrollo de software, como software libre y *copyleft*, que se han de adoptar para que la propuesta pueda superar algunos impedimentos de orden legal, relacionados con la propiedad intelectual de textos originales y traducciones.

El problema de la equivalencia

Una de las premisas más firmes en la historia de la traducción automática ha sido considerar que la traducción es fundamentalmente un problema de equivalencia semántica. Esta premisa se asienta en el supuesto, que se remonta a Leibniz, y que recogieron Frege y Montague, padres de la semántica contemporánea, de que todas las lenguas del mundo comparten una misma subestructura lógica. Se sigue así que si fuéramos capaces de descubrir y formalizar esta subestructura, el problema de la traducción estaría resuelto.

Movida por este razonamiento, durante varias décadas la comunidad científica internacional ha centrado su atención en el problema de la equivalencia conceptual, bien a través de representaciones neutras y comunes - técnica de interlingua- o proyectando representaciones intermedias entre pares de lenguas - técnica de transferencia. Entre los modelos más utilizados para el tratamiento computacional de la semántica cabe destacar los siguientes: redes semánticas (Simmons y Slocum, 1972), preferencias semánticas (Wilks, 1973), gramáticas de caso y valencias (Somers, 1987), representaciones conceptuales (Carbonell et al, 1981; Nirenburg et al, 1985), transferencia léxica (Melby, 1988; Alonso,1990), semántica léxica (Dorr, 1993) y desambiguación léxica (Masterman, 1957; Amsler y White, 1979).

Estas citas son sólo una pequeña muestra de una vastísima producción científica que, aunque ha tenido aplicación en otras áreas de la lingüística computacional, ha sido en la traducción automática donde se ha probado de manera más intensa. Sin embargo, pese a las cotas de excelencia alcanzadas en el plano teórico, los resultados prácticos de los sistemas diseñados han sido insatisfactorios. Es una situación que sólo unos pocos observadores autorizados dentro del colectivo científico, como Melby (1995) o Kay (1997), se han atrevido a señalar, a modo de crítica velada hacia sus propios colegas. Con distintos matices, ambos autores coinciden en lo inapropiado de la metodología empleada, pero ha sido Melby quien de manera más explícita ha cuestionado la hipótesis de la universalidad conceptual entre las lenguas. Melby duda de la existencia de unidades conceptuales universales, comunes a todas las lenguas, y advierte de lo utópico de este método para la traducción automática.

Los traductores profesionales han dudado siempre de estos métodos, como queda reflejado en algunos populares foros de Internet (Lantra-l). En el campo de la traductología, además, existen estudios recientes que describen otros niveles de equivalencia de no menor importancia que el semántico. Nord (1993), como autora más destacada en el estudio de la equivalencia en traducción, propone dos dimensiones más, la equivalencia estilística y la equivalencia pragmática. Por otro lado, Hatim y Mason (1990), insisten en la importancia de considerar la traducción una cuestión de índole sobre todo pragmática, más que meramente lingüística, y proponen un nivel más abstracto de equivalencia, en el plano de los símbolos sociales y culturales, esto es, de la semiótica.

La unidad de traducción

Este debate sobre la equivalencia nos introduce de lleno en otra cuestión polémica, la unidad de traducción. Si se mantiene que traducir consiste fundamentalmente en relacionar representaciones semánticas de textos en distintas lenguas, parece obvio que la unidad de traducción debería tener una dimensión conceptual. Durante años así se ha considerado, como refleja la bibliografía especializada (Bennett, 1994). Por el lado de la traducción humana, Vinay y Darbelnet (1958) y Vázquez Ayora (1977) son dos referencias obligadas que ya incluían - con otros nombres- patrones de subcategorización, construcciones colocativas, lexías complejas y giros idiomáticos como unidades. La definición de Vinay y Darbelnet establece la unidad como "el menor segmento del enunciado en el que la cohesión de los signos es tal que no se entenderían si fueran traducidos por separado". Es decir, equipara la unidad de traducción con la unidad de significado, que a su vez se corresponde con la unidad lexicológica. Pero este enfoque es limitado y no puede dar cuenta de unidades mayores o más complejas, ni de dimensiones distintas de la semántica.

Intentando poner un poco de orden en la variedad, suelo sugerir a mis alumnos esta clasificación de unidades de traducción:

a. **Categorías morfosintácticas:** la unidad básica en todos los sistemas de traducción automática suele ser la palabra (o lexía simple). Las categorías morfosintácticas permiten establecer abstracciones sobre las palabras (*el/the* > Det; *eye/ojo* > N; *happy/feliz* > A; *eat/comer* > V; *over/sobre* > P) y son la base de las gramáticas de estructura sintagmática: SN = Det N.
b. **Subcategorías:** dentro de cada categoría se da una gran variedad de comportamientos, la mayoría divergentes entre una lengua y otra. Los patrones de subcategorización permiten plasmar estas divergencias: subj(x) *likes* obj(y) / subj(y) *gusta* obj(x).
c. **Colocaciones:** categorías y subcategorías muestran con frecuencia "hábitos de colocación sintagmática" particulares: *fast waltz*, *rapid movement*, *quick action*, *speedy recovery*.
d. **Lexías complejas** (palabras compuestas): combinaciones de palabras que lexicalizan: *comida rápida/fast food*; *movimientos oculares rápidos/apid eye movement* (*REM*).
e. **Locuciones:** grupos preposicionales o conjuntivos fijos: *after all/när allt kommer omkring, a pesar de todo*

omkring, *still*/*a pesar de todo*.

f. **Giros idiomáticos:** son grupos sintagmáticos con flexibilidad sintáctica: *Estaba más loca que una cabra*/*She was as nutty as a fruitcake*.

g. **Fórmulas:** incluye proverbios, *Tanto monta, monta tanto, Isabel como Fernando*; títulos de obras, películas *Monthy Pyton and the Holy Grail*/*Los caballeros de la mesa cuadrada*; y otros elementos fijos del discurso, como este extracto de una escritura inglesa *To do all such other things as are incidental or conductive to the above objects or any of them*.

Ante esta clasificación surgen varias cuestiones. La primera es dilucidar si todas las unidades propuestas pueden ser recogidas en los diccionarios tradicionales, dado que, al menos desde el enfoque semántico, es en ellos en los que recae la función de establecer equivalencias. El problema es que las fórmulas, por su tamaño y variedad, hacen la tarea impracticable; los giros idiomáticos suelen estar por lo general pobremente representados y, en cualquier caso, no existe diccionario bilingüe conocido que recoja toda la información lexicológica necesaria en traducción de manera sistemática, ni homogénea. De entre la multitud de diccionarios bilingües que existen, algunos contienen información de subcategorización, de colocaciones más frecuentes, de lexías complejas (sobre todo si son especializados), o de locuciones y giros idiomáticos, pero ninguno es exhaustivo. El problema de la exhaustividad en los diccionarios es un problema antiguo e insoluble. Los diccionarios en su concepción son depósitos estáticos, que requieren una laboriosa labor de compilación y validación, frente a la formación de nuevas palabras, giros y otras expresiones, que está siempre activa y es dinámica.

Otra cuestión es la composicionalidad. En el enfoque puramente semántico de la traducción la noción de composicionalidad desempeña un importante papel. Se dice que la traducción de un texto debe ser un proceso composicional, en el sentido de que la traducción de una expresión compleja es una función de la traducción de sus partes constituyentes. En la clasificación de unidades de traducción mencionada se refleja un continuo entre unidades simples y complejas. Internamente las más complejas no son composicionales y por eso deben tratarse como unidades. Las colocaciones ocupan un lugar intermedio, algunos autores sostienen que son composicionales (Pustejovsky 1993, Viegas et al 1998), lo que en teoría permitiría tratarlas de manera eficiente en los diccionarios. En la práctica sin embargo distan de estar convenientemente contempladas. Más problemático es el tratamiento de las fórmulas, que son fundamentalmente unidades semióticas y, de acuerdo con la tesis de Hatim y Mason (1990), no se someten a las reglas de la semántica ni entran en el juego de la composicionalidad. Por ello, un sistema de traducción tiene que resolver el problema de la equivalencia atacando primero la identificación de unidades por el lado de las unidades no composicionales y más complejas, y sólo recurrir a las simples después.

Corpora multilingües en TMX

Una alternativa a los diccionarios como fuente única de información son los corpora multilingües. Estos son colecciones de textos en distintos idiomas, cuyo valor se multiplica si son debidamente procesados y anotados. Para un par determinado de lenguas, si el corpus es suficientemente grande y representativo, la información que aporta puede ser tan completa o más que la del mejor diccionario. La disponibilidad creciente de texto en formato electrónico hace relativamente fácil la labor de compilar corpora y se ha avanzado mucho en el tratamiento computacional (Abaitua, 2000). Si los corpora son paralelos, es posible obtener porcentajes cercanos al 100% para la alineación tanto de palabras como de oraciones (Catizone et al. 1989; Gale y Church, 1993; Kay y Rscheisen, 1993; Martínez, 1999).

Un corpus alineado y anotado constituye una *memoria de traducción*. Las memorias de traducción (MMT) son una tecnología alternativa a la traducción de base semántica y tienen su origen en una propuesta de Nagao (1984) llamada traducción "por ejemplos". Los sistemas que utilizan esta tecnología no traducen mediante reglas que equiparan representaciones conceptuales, sino mediante analogías o comparaciones entre el texto que se desea traducir y los ya traducidos almacenados en la memoria. Son muy adecuados para textos que contengan un alto porcentaje de expresiones formulaicas y giros idiomáticos, como es el caso de los textos de especialidad. No sirven para textos creativos o expresivos, para los que de todas formas tampoco dan buenos resultados los métodos basados en reglas y requieren traducción humana.

En ámbitos de textos repetitivos, como son los manuales de uso y referencia, los documentos administrativos, los partes informativos (bolsa, meteorología, sucesos), pero sobre todo en el ámbito de la traducción y adaptación de productos de software (localización), las memorias de traducción suponen una interesante opción. Esto se ha reflejado en el mercado de software, que ha visto incrementar de forma significativa el número de ofertas de sistemas comerciales: Déjà-Vu (ATRIL), Translator's Workbench (TRADOS), Transit (STAR), SDLX, etc.. Grandes empresas, instituciones y muchas agencias de traducción han adquirido alguno de estos sistemas para mecanizar en parte sus proyectos de traducción y localización.

Pero los sistemas MMT tienen un inconveniente y es que antes de ser productivos y rentables, antes de que empiecen a ofrecer resultados operativos, precisan un laborioso proceso de alimentación, es decir, de construcción y optimización de la memoria. Esta tarea puede requerir considerables dosis de dedicación y esfuerzo. Un segundo problema, derivado en parte del anterior, es la dependencia del software utilizado. El coste de adquisición de los sistemas MMT es muy alto y su puesta a punto muy costosa, así que en consecuencia, es muy complicado migrar de un sistema a otro. Para paliar este inconveniente es para lo que se ha diseñado el formato TMX (translation memory exchange format; Melby, 1998). En la actualidad, la mayoría de los sistemas MMT disponen de filtros de importación y exportación a TMX.

El formato TMX está basado en el metalenguaje XML y consta de una colección sencilla de etiquetas para marcar los elementos básicos de una memoria de traducción. Es en este sentido una alternativa a otras propuestas de etiquetado conocidas, como puede ser fundamentalmente TEI (Erjavec, 1997). Como TEI es un modelo de etiquetado más genérico y también más rico, no resulta complicado pasar de un corpus TEI a una memoria en TMX.

<TU>
<TUV lang="EN" creationdate="1600" creationid="William Shakespeare" changedate="1951" changeid="Peter Alexander/Collins">
<SEG>
Hamlet
The Scene: Denmark.
Act One
Secene I. Elsinore. The guard-platform of the Castle. Francisco at his post. Enter to him Bernardo
[...]
Exeunt marching. A peal of ordance shot off.</SEG></TUV>
<TUV lang="ES" creationdate="1929" creationid="Luis Astrana Marín/Aguilar" >
<SEG>
Hamlet, príncipe de Dinamarca
Escena: Elsinor
Acto primero
Escena I.- Elsinor.- Explanada delante del castillo
Francisco, de centinela en su puesto.- Entra Bernardo dirigiéndose a él
[...]
Marcha fúnebre. Salen, llevándose los cadáveres. Después se oye una descarga de artillería.</SEG></TUV>
<TUV lang="ES" creationdate="1994" creationid="José María Valverde/Planeta">
<SEG>
Hamlet
La acción, en Elsinor
Acto primero
Escena Primera
Elsinor. Ante el castillo
Entran Bernardo y Francisco, centinelas
[...]
Se van marchando; después, se disparan salvas de artillería.</SEG></TUV>
<TUV lang="ES" creationdate="1994" creationid="Ángel-Luis Pujante/Espasa">
<SEG>
La tragedia de Hamlet, príncipe de Dinamarca
I.i Entran Bernardo y Francisco, dos centinelas
[...]
Salen en marcha solemne, seguida de una salva de cañón.</SEG></TUV>
</TU>

<TU>
<TUV lang="EN" creationdate="1600" creationid="William Shakespeare" changedate="1951" changeid="Peter Alexander/Collins">
<SEG>Exeunt marching. A peal of ordance shot off.</SEG></TUV>
<TUV lang="ES" creationdate="1929" creationid="Luis Astrana Marín/Aguilar" >
<SEG>Marcha fúnebre. Salen, llevándose los cadáveres. Después se oye una descarga de artillería.</SEG></TUV>
<TUV lang="ES" creationdate="1994" creationid="José María Valverde/Planeta">
<SEG>Se van marchando; después, se disparan salvas de artillería.</SEG></TUV>
<TUV lang="ES" creationdate="1994" creationid="Ángel-Luis Pujante/Espasa">
<SEG>Salen en marcha solemne, seguida de una salva de cañón.</SEG></TUV>
</TU>

Tabla 1. Traducciones de Hamlet: ejemplos en TMX

En TMX la definición de una unidad de traducción es muy simple: cualquier cadena de caracteres entre las etiquetas <TU>...</TU>. Una <TU> puede estar formada por tantas variedades lingüísticas o estilísticas <TUV> como sean necesarias, cada una de ellas, debidamente documentada (Tabla 1). El tamaño de la unidad de traducción no está limitado, así que nada impide que toda una obra literaria pueda ser tratada como unidad

Octubre 2001