DISEÑO Y USO DE UN CORPUS PARALELO HOLANDÉS-ESPAÑOL

JOS HALLEBEEK

Universidad de Nimega

1. EL CORPUS EN LA LINGÜÍSTICA ACTUAL

En la lingüística moderna el corpus constituye la base para una forma de lingüística empírica. En el corpus se hallan los datos que el lingüista investiga y analiza para llegar a formular o comprobar sus hipótesis sobre la lengua. Estas hipótesis pueden referir al sistema de la lengua (la competencia) pero más aún tienen que ver con el uso de la lengua (la actuación) en vista de que el corpus es una abundante base de datos de hechos lingüísticos concretos. El corpus forma una colección de textos que reúne una o más variedades de determinada lengua (o determinadas lenguas), seleccionadas según una serie de criterios previamente establecidos.

1.1. Corpus bilingües, multilingües y paralelos

El corpus no contiene necesariamente textos en una sola lengua. Puede ser de dos (corpus bilingüe) o de más lenguas (corpus multilingüe). En tales casos los textos del corpus no son textos reunidos arbitrariamente, sino que están escogidos según idénticos criterios de selección en una y otra lengua. Por ejemplo, el *Aarhus* Corpus *of Danish, French and English* está compuesto de textos en tres lenguas que tratan todos del mismo tema: el derecho de contrato. No son traducciones de los mismos textos. Los textos son diferentes pero coinciden en la temática.

Cuando un corpus tiene los mismos textos en diferentes lenguas se habla de un *parallel corpus*: corpus paralelo. Un ejemplo de un corpus paralelo es el *Hansard* Corpus, que contiene una colección de actas del parlamento canadiense tanto en versión inglesa como en la francesa. El corpus paralelo se usa para la construcción automática de léxicos y para la investigación sobre la traducción. Para poder manejar estos corpus es necesario que las frases y las palabras que son traducciones mutuas sean alineadas, es decir puestas unas al lado de otras.

En la universidad de Oslo Johansson y Hofland[1] están compilando un corpus de un millón de palabras que consiste de 34 parejas de texto en inglés y en noruego. Son fragmentos de 10.000 o 15.000 palabras sacados de obras originales y de su traducción

[1]Stig Johansson y Knut Hofland: "Towards an English-Norwegian parallel corpus", en Fries *et al.*, 1994, pp. 25-37; y Stig Johansson y Jarle Ebeling: "Exploring the English-Norwegian Parallel Corpus", en Carol E. Percy, Charles F. Meyer e Ian Lancashire (eds.): *Synchronic* corpus *linguistics*, Amsterdam-Atlanta, Rodopi, 1996, pp. 3-15.

Lengua y Cultura. HALLEBEEK, Jos. Diseño y uso de un corpus paralelo holandés-español
Centro Virtual Cervantes

al inglés o al noruego. Es decir, que se parte de originales en las dos lenguas. No dan los datos exactos de los textos incluidos en el corpus. Se limitan a mencionar dos novelas noruegas y dos inglesas, el texto del acuerdo económico europeo, y del acuerdo de Maastricht (el último aún no en su poder), y posibles textos de la agencia noruega de prensa. Como el corpus en inglés no sólo contiene textos originales sino también textos traducidos del noruego, se trata al mismo tiempo de un corpus comparable: este nombre se reserva para corpus que tiene textos originales y traducidos en la misma lengua. Por otra parte los corpus de los originales en inglés y en noruego están diseñados según los mismos criterios de selección, y en este sentido la combinación de los dos constituye un corpus bilingüe. Los textos que forman parte del corpus de un millón de palabras estarán alineados. En la parte del corpus alineada hasta la publicación de Johansson y Ebeling[2] se han puesto una al lado de otra las frases (los enunciados) de los textos, no las palabras sueltas. El equipo tiene otros textos en preparación para agregar al corpus en algún momento del futuro. Los textos están codificados según las normas TEI (*Text Encoding Initiative*). Puesto que las unidades de los textos estarán identificadas detalladamente y los originales con sus traducciones estarán alineados, un programa de búsqueda producirá parejas de enunciados (frases) para cualquier palabra o expresión en cada una de las dos lenguas.

El corpus paralelo inglés-sueco está proyectado para incluir una variedad de géneros de texto según el modelo del LOB y del Brown corpus. Incluye textos originales en las dos lenguas con sus traducciones, en fragmentos de 10.000 a 15.000 palabras. El corpus está en vías de compilación.

El Chemnitz corpus de traducción inglés-alemán es un corpus paralelo de textos originales tanto en inglés como en alemán junto con sus traducciones. Tendrá las siguientes categorías de texto, mayormente en fragmentos de 20.000 palabras:

— publicaciones de la comunidad europea en terreno de la economía y de la política social;

— libros de texto científicos: historia, filosofía, ciencias políticas, económicas y naturales;

— folletos turísticos;

— publicaciones de la Embajada Británica de Bonn;

— literatura contemporánea.

El número total de palabras será de 1.500.000. Los textos literarios ocupan sólo una parte marginal del corpus y no forman parte del *core corpus*.

1.2. Procedimientos utilizados en la alineación de palabras y frases

Para llegar a alinear automáticamente las oraciones y las palabras de textos en lenguas diferentes se utilizan en primer lugar métodos puramente estadísticos. Así es, p. ej., el método Gale y Church,[3] desarrollado con ayuda del *Hansard* Corpus, que se basa en dos criterios:

a) oraciones más largas en una lengua se traducen en secuencias más largas en otra;

---

[2] Art. cit.

[3] W. A. Gale y K. W. Church: "A program for aligning sentences in bilingual corpora", *Computational Linguistics* (1993), 19, 1, pp.75-102

Lengua y Cultura. HALLEBEEK, Jos. Diseño y uso de un corpus paralelo holandés-español
Centro Virtual Cervantes

b) unos tipos de alineamiento se hallan con más frecuencia que otros: es más probable que una frase en una lengua se traduce también por una frase en la otra.

De modo que una frase traducida por dos, o dos frases traducidas por una son analogías menos frecuentes.

Para mejorar los sistemas de alineamiento se han ido introduciendo en el reconocimiento de traducciones los llamados *cognates*, que no son sólo palabras sino más bien signos comunes a las dos lenguas. Se llaman *anchor points* (puntos ancla), o sea puntos de correspondencia conocida entre dos lenguas. Entre inglés y francés, por ejemplo, tenemos los siguientes: los signos de interrogación y de exclamación; las palabras que tienen cierto número de caracteres idéntico al principio de la palabra (*tax* y *taxe*); los nombres propios; las expresiones numéricas; los paréntesis. Se están llevando a cabo experimentos para comprobar la similitud entre dos palabras de diferentes lenguas, que no coinciden por completo. Hasta qué punto coinciden los caracteres y cuántas modificaciones son necesarias para convertir una palabra de una lengua en su equivalente en otra: por ejemplo, *couleur* y *colour* necesitan 2 cambios.

1.3. Aplicaciones de corpus en la investigación sobre la traducción

Los corpus paralelos bilingües se usan en la traducción automática en sistemas de traducción estadística basada en alineamiento léxico y la posición de palabras. Suelen tomarse en consideración no frases enteras sino secuencias de tres palabras, para las que se busca la equivalencia en otra lengua. Los corpus son una fuente de datos directa para las máquinas. A base del principio de la analogía sacan del corpus ejemplos típicos de frases o partes de ella para llegar a realizar la traducción de textos no traducidos todavía. Las últimas tendencias en los sistemas de traducción automática se alejan cada vez más de análisis sintácticos y semánticos completos utilizando gramáticas de reglas formales para ir basándose en datos de uso de la lengua viva.

En *Example-Based Machine Translation (EBMT)* la traducción se realiza según el siguiente procedimiento. Se dispone de un corpus bilingüe de textos traducidos y alineados. Al ofrecerse una frase para traducir, el programa busca en el corpus si ya está la misma frase alineada con su traducción en la otra lengua. La probabilidad de encontrar la misma frase ya traducida no es muy alta, a no ser que se disponga de un corpus de muchos millones de frases traducidas y alineadas. Por esto, también se ha ideado un sistema que no busca la traducción de la frase entera sino de partes de ella. Usando gramáticas sintácticas, tanto en la lengua fuente como en la lengua objeto se llega a dividir las frases en sus constituyentes funcionales. Luego se buscan las equivalencias de esas partes en una y la otra lengua. La ventaja de este sistema es que las frases completas no necesitan ser idénticas y que se pueden aprovechar partes de diferentes frases. El sistema no ha sido probado todavía.

El corpus paralelo que contiene textos originales con su traducción en otra u otras lenguas se utiliza en la formación de traductores y para mejorar los resultados de sistemas de traducción automática. En el primer caso, su contribución esencial es que contiene evidencia de cómo un traductor resuelve problemas que se ofrecen en la práctica de la traducción. Al comparar un corpus de textos originales sobre determinado tema con textos traducidos sobre el mismo tema y en la misma lengua pueden compro-

Lengua y Cultura. HALLEBEEK, Jos. Diseño y uso de un corpus paralelo holandés-español
Centro Virtual Cervantes

barse ciertas propiedades características. Por ejemplo, que el número de tipos de palabras (*types*) es inferior en las traducciones, que la relación palabras gramaticales vs. palabras léxicas es más baja en los textos traducidos, o que determinadas estructuras oracionales son más típicas para una u otra clase.

1.3.1. Estudios realizados

Hasta el momento se ha realizado un estudio sobre las interrogaciones en inglés y en noruego en Wikberg[4] utilizando el corpus paralelo. Wikberg comparó textos procedentes de novelas: originales en inglés con su traducción al noruego, así como textos originales en noruego con su traducción al inglés. Sacó del corpus las oraciones de 7 a 10 palabras para obtener oraciones finitas y no fragmentos de oraciones. El signo de interrogación resultó ser muy útil para encontrar las interrogaciones en el texto original y traducido. Oraciones interrogativas no son siempre simples preguntas sino que expresan también ruegos, sugerencias, etc. Ha encontrado contrastes formales, semánticos y pragmáticos entre las dos lenguas. A veces una interrogación en una lengua se convierte en una oración declarativa en la otra. Además el autor ha podido establecer diferencias en la distribución de estructuras sintácticas equivalentes en los textos originales y sus traducciones.

Aprovechando el corpus paralelo inglés-sueco, M. Johansson estudia el fenómeno de *fronting* en sueco y en inglés partiendo de tres obras suecas y tres obras inglesas con sus correspondientes traducciones. Los seis textos representan tres géneros: ficción general, novela policíaca y autobiografía. Para Johansson *fronting* refiere a la colocación de cualquier elemento delante del sujeto (en una oración declarativa) en inglés y delante del verbo + sujeto en sueco, con excepción de conjunciones, pronombres relativos e interrogativos, interjecciones como *well, oh,* etc. Las diferencias y coincidencias existentes entre ambas lenguas en cuanto al fenómeno en cuestión aparece con más claridad en la comparación de textos originales con su traducción. En inglés el fenómeno de *fronting* es mucho menos frecuente que en sueco en el que el objeto directo o un complemento adverbial aparece muchas veces al principio de la oración. Esto ya se sabía por intuición. Ahora se ve confirmada esta intuición en la comparación de textos concretos. Johansson estudia detalladamente las clases de adverbiales y de otras funciones sintácticas que se hallan en posición inicial antepuestas al sujeto o al verbo + sujeto. En realidad hay cuatro posibilidades:

— un elemento ocupa la misma posición en sueco y en inglés;

— un elemento cambia de posición;

— se usa una estructura diferente (por ejemplo, un adverbio es sustituido por una proposición subordinada);

— se suprime el elemento en la traducción.

[4] Kay Wikberg: "Questions in English and Norwegian: Evidence from the English-Norwegian Parallel corpus", en Percy y otros, o. cit., pp. 17-28.

Lengua y Cultura. HALLEBEEK, Jos. Diseño y uso de un corpus paralelo holandés-español
Centro Virtual Cervantes

Utilizando el Chemnitz corpus de traducción inglés-alemán Schmied y Schäffler[5] dedican atención al fenómeno de lo que en inglés se llama *translationese*: hablando en términos generales, refiere este fenómeno a desviaciones del uso corriente de la lengua presentes en textos traducidos y causadas por influencia del texto original (fuente). Las desviaciones pueden ser simplemente faltas contra el sistema de la lengua destino (por ejemplo no usar en traducción española un subjuntivo porque no se usa en la lengua fuente). También puede que la traducción vaya en contra de la norma de uso establecida en determinada lengua, por ejemplo, usar en traducción holandesa una proposición *cleft* (hendida) que en español se usa con frecuencia pero que en holandés no es corriente aunque tampoco es imposible (o sea un error gramatical). Existen también características universales de traducción que no se deben a influencias de determinada lengua fuente, como la tendencia a simplificar oraciones, a desambiguar elementos, a evitar repeticiones.[6] Hasta ahora existen pocos estudios sobre la norma de uso de lenguas específicas, sea inglés, alemán o español. Sólo al saber cómo es la norma podemos constatar desviaciones de la misma. El corpus *comparable* puede servir para llegar a definir esa norma. Si hay algún conocimiento suelto, está basado en estudios contrastivos o intuición sobre propiedades de lenguas. Así sabemos que en inglés se usan verbos modales (*may*, etc.) donde el alemán o el holandés emplean adverbios modales. En español el empleo de los pronombres personales tónicos de sujeto es mucho más limitado que en inglés, alemán u holandés. ¿Es un fenómeno que pertenece a la norma y al sistema? Hay muchos factores que juntos juegan un papel en el análisis y la comparación de textos traducidos: gramaticalidad, aceptabilidad (sí o no según la norma), contexto sociocultural, interpretación correcta del significado. En su artículo Schmied y Schäffler estudian si es cierta la afirmación de que el inglés es menos directo que el alemán. Tratan de encontrar diferencias cuantitativas y cualitativas en los elementos y estructuras que expresan lo tentativo y lo indirecto: adverbios y verbos modales, construcciones impersonales, comentarios parentéticos, etc. Tomando en cuenta textos originales en inglés y en alemán así como traducciones de los textos ingleses al alemán hacen un análisis pormenorizado de los elementos indicados. Ese análisis es complicado y se corre el riesgo de incurrir en interpretaciones subjetivas.

2. HACIA UN CORPUS PARALELO HOLANDÉS-ESPAÑOL

Viendo las experiencias y los resultados provisionales obtenidos en la confección y el análisis de corpus paralelos nos parece interesante empezar un experimento con un corpus de este tipo que contiene textos originales en holandés y en español acompañados de sus traducciones a las respectivas lenguas. Una vez alineados los textos las aplicaciones para la investigación y la enseñanza de idiomas son múltiples. Por lo que respecta a esto último pensamos sobre todo en la formación de traductores y en los estudios de gramática descriptiva y contrastiva de las dos lenguas.

---

[5] Josef Schmied y Hildegard Schäffler: "Approaching translationese through parallel and translation corpora", en Percy y otros, o. cit., pp. 41-56.

[6] Mona Baker: "Corpus linguistics and translation studies. Implications and applications", en M. Baker y otros, 1993, p. 243.

Lengua y Cultura. HALLEBEEK, Jos. Diseño y uso de un corpus paralelo holandés-español
Centro Virtual Cervantes

No abundan las traducciones de textos no literarios de índole general entre el español y el holandés. Nos referimos a manuales de historia, ciencias sociales, antropología, etc. Por otra parte, los textos que suelen producir las oficinas de la Unión Europea, muchas veces disponibles en las lenguas de la comunidad, son de carácter técnico y van destinados a especialistas en agricultura, política, economía, etc. Esto los hace poco interesantes para alumnos universitarios cuyos conocimientos de la segunda lengua que estudian no permite aún el manejo de textos de una complicación léxica de este tipo.

Se ha demostrado que las traducciones de folletos turísticos no cumplen siempre con las exigencias de calidad que se les debería imponer. Aunque sí es cierto que representan un tipo de textos que en sí es interesante analizar por su carácter persuasivo.

Antes de decidir por un corpus extenso que contenga diversas variedades de géneros de textos, optamos por la formación de un corpus de tamaño limitado compuesto de textos literarios tomados de novelas de autores contemporáneos. En general, actualmente la calidad de las traducciones de este tipo de obras es bastante buena. La fantasía y el dominio del lenguaje de los novelistas se refleja en los textos traducidos, interesantes desde el punto de vista del estilo y de los recursos sintácticos, morfológicos y léxicos empleados.

Nos proponemos por lo tanto desarrollar un corpus paralelo de 8 fragmentos de 20.000 palabras cada uno, procedentes de 8 novelas originales españolas y 8 novelas originales holandesas, con sus correspondientes traducciones. Este corpus tendrá el carácter de un corpus piloto del volumen de unas 640.000 palabras en total.

2.2. Marcación de los textos del corpus

Para una detallada identificación de los textos del corpus y de sus partes vamos a agregar marcas (etiquetas) con información sobre:

— si es texto original o traducido
— el autor
— el título de la obra
— el capítulo
— el párrafo
— la página
— la línea
— el número de la frase.

Las frases son las unidades independientes en que se dividen los textos, reconocibles ortográficamente por la mayúscula al principio y el punto final. Por regla general un punto final seguido por una letra mayúscula marca el principio de la frase. Hace falta calcular el número total de las frases con vistas a la comparación de texto original y texto traducido, así como la correcta alineación de las frases.

2.3. Alineación de las frases

Preferimos empezar con una alineación sólo de las frases del original y de su traducción. Por experiencia propia sabemos que cuando se trata de textos literarios los

Lengua y Cultura. HALLEBEEK, Jos. Diseño y uso de un corpus paralelo holandés-español
Centro Virtual Cervantes

traductores no mantienen a rajatabla la misma repartición del texto en frases. En español, éstas suelen ser más largas que en holandés. Por esto fácilmente el traductor las corta en unidades independientes más pequeñas.

Si ya será complicado llegar a una alineación exacta de las frases entre original y traducción, mucho más difícil aún resultará llevar a cabo una alineación automática de las palabras o de los sintagmas. Sólo sería posible si se tomara como base el significado. Pero es bien conocido que hasta el momento falta un sistema que abarque la descripción de los aspectos semánticos de palabras, grupos de palabras y frases (oraciones). Traducir no consiste en ir sustituyendo palabra por palabra los elementos léxicos de una lengua por los correspondientes elementos de otra lengua, conservando la categoría de cada una de las palabras. Basta con señalar algunas diferencias esenciales. Hay verbos españoles que se traducen por adverbios en holandés (*volver a, acabar de, gustar,* etc.), adverbios en español que se convierten en afijos en holandés (el comparativo y el superlativo), artículos españoles que quedan sin traducción en holandés, y pronombres personales del holandés que no tienen equivalentes en la traducción al español. Y no hablamos de los cambios en el orden de las palabras que se producen entre ambas lenguas. También hay un empleo corriente de partículas en holandés como *maar, toch, wel* que en muchos casos quedan sin traducción en español. Está claro que una tentativa de alinear automáticamente textos originales y traducidos en holandés y en español palabra por palabra sin tomar en cuenta la significación sólo dará lugar a una larga serie de errores. El tiempo que habría que invertir para eliminar esos errores a mano sería muy largo.

Para alinear las frases de original y traducción nos proponemos utilizar:

— una lista de palabras ancla del español y del holandés a la Johansson, que deberá incluir las palabras funcionales, bases de las palabras léxicas más frecuentes y los nombres propios usados en el original;

— junto con otra lista de elementos no léxicos como los signos de puntuación, marcas de capítulos, párrafos, frases.

2.4. Exploración del corpus

El corpus nos permitirá hacer investigaciones sobre:

— el lenguaje usado en los textos originales del holandés o del español;

— el lenguaje usado en los textos traducidos al holandés o al español;

— la relación entre original en una lengua y su traducción a la otra.

Sobre todo esta última perspectiva nos interesa para poder detectar y describir problemas de traducción entre las dos lenguas situándonos en el terreno de la lingüística contrastiva y descriptiva. Los temas que investigar son múltiples. Además de los que hemos citado en las investigaciones revisadas en la primera parte de este artículo pensamos concretamente —para citar algunas posibilidades— en estudios sobre diferencias del orden de palabras entre español y holandés, sobre la frecuencia y las reglas de uso de la construcción pasiva y de la oración coordinada y subordinada, sobre el uso de las formas no personales del verbo, sobre la traducción del verbo copulativo *worden* del holandés al español, sobre la traducción de partículas del tipo *wel, toch* y otros que no hace falta citar.

Lengua y Cultura. HALLEBEEK, Jos. Diseño y uso de un corpus paralelo holandés-español
Centro Virtual Cervantes

Además nos parece interesante ver cómo podemos utilizar el material del corpus para desarrollar un programa de entrenamiento de traductores del holandés al español y del español al holandés. Este programa tendrá la forma de material de enseñanza asistida por ordenador. En él se tratarán problemas gramaticales generales de la traducción entre las dos lenguas y también problemas específicos de la traducción literaria.

BIBLIOGRAFÍA

BAKER, Mona: "Corpora in translation studies. An overview and some suggestions for future research", *Target* (1995), 7.2, pp. 223-243.

JOHANSSON, Mats: "Fronting in English and Swedish: A text-based contrastive analysis", en Percy y otros, 1996, pp. 29-39.

MCENERY, Tony y Michael OAKES: "Sentence and word alignment in the CRATER Project", en THOMAS y otros, 1996, pp. 211-231.

SÁNCHEZ LEÓN, F.: "Desarrollo de un etiquetador morfosintáctico para el español", *Procesamiento del Lenguaje Natural*, 17, sept. de 1995, pp. 14-28.

THOMAS, Jenny y Mick SHORT (eds.): *Using Corpora for Language Research*. Studies in Honour of Geoffrey Leech, 1996.

TSUJII, J., S. ANANIADOU, J. CARROLL y S. SEKINE: *Methodologies for the Development of Sublanguage MT System II*, CCL UMIST Report, 91/11, 1991.

WILSON, A. y A. MCENERY (eds.): *Corpora in Language Education and Research: A Selection of Papers from Talc94*, Unit for Computer Research on the English Language, Technical Papers 4 (special issue) Lancaster University, 1994.

ZANETTIN, F.: "Parallel words: designing a bilingual database for translator activities", en Wilson y otros, 1994, pp. 99-111.

Lengua y Cultura. HALLEBEEK, Jos. Diseño y uso de un corpus paralelo holandés-español
Centro Virtual Cervantes