CORPUS MULTILINGÜES PARA LA INVESTIGACIÓN Y LA ENSEÑANZA DE LA TRADUCCIÓN: EL PROYECTO MUST

Nava Maroto García
Universidad Politécnica de Madrid
Arsenio Andrades Moreno
Universidad de Cádiz

RESUMEN

Se exponen las características fundamentales del proyecto MUST (MUltilingual Student Translation), sus principales objetivos, el proceso de recopilación de textos tanto a escala global como los que se aportan específicamente desde el equipo español de la Universidad Complutense de Madrid. Asimismo, se presenta la interfaz de gestión del corpus basada en la red (Hypal4MUST) que incorpora funciones esenciales para el análisis textual y lingüístico, como son el etiquetado POS, alineación automática, anotación y búsquedas (Obrusnik, 2014). Finalmente, damos a conocer ejemplos textuales recopilados en la UCM e información cualitativa sobre los estudiantes involucrados en el proceso.

Palabras clave: proyecto MUST, corpus, traducción, etiquetado

ABSTRACT

This article presents MUST (MUltilingual Student Translation) project's main features, its essential goals, the text compilation process both from a global and a local perspective, focusing particularly on the work done by the Spanish team at the Universidad Complutense de Madrid. The web-based Hypal4MUST interface, which incorporates POS tagging, automatic sentence alignment, annotation and corpus search features (Obrusnik, 2014). Finally, some examples of the texts collected at UCM are presented and qualitative information about the students involved in the process is provided.

Keywords: MUST project, corpus, translation, tagging

1. INTRODUCCIÓN

La investigación en el área de los Estudios de Traducción ha experimentado un impulso considerable en los últimos años (Granger y Lefer, 2017), que se ha visto reforzado significativamente gracias al uso de la Lingüística de Corpus (LC) como metodología para el análisis de textos traducidos, sobre todo en lo que respecta a corpus de producción de estudiantes: transferencia lingüística, análisis de errores y estrategias aplicables a la enseñanza, aprendizaje y adquisición de idiomas y de traducción (Johansson, 2007; Rica et al., 2014; Rica y Braga, 2015; Rica, 2018). La LC, por lo tanto, se ha convertido en una herramienta metodológica clave para el análisis lingüístico textual, favorecida igualmente por la proliferación de programas de software y recursos electrónicos.

Presentamos el proyecto MUST (*Multilingual Student Translation*), una iniciativa de la Universidad Católica de Lovaina, que aúna la práctica de la traducción y la lingüística de corpus, y en la que participan un gran número de universidades a nivel internacional. El corpus MUST incorpora traducciones de tipología variada realizadas por estudiantes universitarios en distintas combinaciones lingüísticas.

En primer lugar, describimos el contexto del proyecto MUST (sección 2). A continuación (sección 3), presentamos la interfaz que estamos utilizando para la compilación, alineación y anotación del corpus, y describimos el proceso, Hypal4MUST. En la sección 4 describimos el sistema de anotación de traducciones MUST-TAS. Más adelante (sección 5) indicamos de manera más detallada cómo contribuye el equipo de la Universidad Complutense de Madrid en el proyecto MUST[1] (participantes, lenguas, tipos de textos y estudiantes implicados). Por último, exponemos algunos ejemplos recogidos en esta primera fase del proyecto (sección 6) y presentamos una recapitulación de los asuntos tratados (sección 7).

[1] El equipo de investigación, que está adscrito al Departamento de Estudios Ingleses: Lingüística y Literatura de la UCM, lo conforman Juan Pedro Rica Peromingo (investigador principal, UCM), Arsenio Andrades Moreno (UCA), Jorge Braga Riera (UCM), Nava Maroto García (UPM), Sara Martínez Portillo (UCM) y Ángela Sáenz Herrero (UNED).

2. EL PROYECTO MUST

MUST es un proyecto internacional impulsado por Sylviane Granger y Marie-Aude Lefer, miembros del *Centre for English Corpus Linguistics* (CECL) de la *Université catholique de Louvain* (UCL), especializado en la compilación y utilización de corpus multilingües con fines pedagógicos y de investigación lingüística.

Los principales objetivos del proyecto MUST consisten en:

- Compilar un amplio corpus de traducciones hechas por estudiantes en distintas combinaciones lingüísticas.
- Desarrollar una interfaz de búsqueda basada en Internet que sea intuitiva para el usuario.
- Diseñar metadatos estándar.
- Crear un sistema de etiquetado de traducciones estándar para las lenguas traducidas.

Las principales características del corpus del proyecto MUST son las siguientes:

- Hasta la fecha están representadas 25 lenguas (alemán, chino, checo, esloveno, español, finés, francés, gallego, griego, inglés, italiano, lituano, macedonio, noruego, polaco, portugués y ruso, entre otros).
- Se incluyen traducciones directas (L2L1) e inversas (L1L2), que pueden ser realizadas por el mismo estudiante.
- No hay restricción en cuanto a tipos textuales, géneros y registros, pero solo se admiten archivos de texto, no audiovisuales.
- Aunque MUST es principalmente un corpus de traducciones de estudiantes, siempre que sea posible se aportarán traducciones de expertos profesionales que sirvan como modelo de referencia. Los participantes pueden ser estudiantes de grado, postgrado o becarios en prácticas.
- Los textos de la lengua de origen y de la lengua de destino incluirán metadatos estándar con información sobre el traductor y las lenguas de origen y destino.

La tipología textual del corpus comprende un amplio espectro de clases de textos y varía en función de la especialización de los investigadores que participan en el proyecto. Los textos de la lengua fuente tendrán una extensión de 250 a 600 palabras.

En una primera fase, el corpus será de acceso y uso exclusivo para los miembros del grupo de investigación, pero el objetivo es que, en una etapa ulterior, los datos recabados sean accesibles a toda la comunidad investigadora para su plena consulta y explotación lingüística.

Se requiere que los estudiantes que aporten sus traducciones tengan un nivel B1 o C2 en la lengua de origen. Todos los participantes firmarán un formulario de consentimiento para que sus traducciones se utilicen con fines de investigación en este proyecto. Los participantes han de facilitar datos personales (sexo, año de nacimiento), información sobre sus lenguas de trabajo y nivel de estudios, así como precisar su competencia lingüística en L1 y L2.

La importancia que concede el corpus del proyecto MUST a los metadatos y al sistema de etiquetado refleja su doble objetivo de cubrir tanto las necesidades de investigación en materia de LC como de traducción.

3. HYPAL4MUST

Para el desarrollo del proyecto (compilación del corpus, alineación, anotación, etc.) se ha desarrollado una herramienta específica accesible a través de la web denominada Hypal4MUST. Esta herramienta está basada a su vez en Hypal (Hybrid Parallel Text Aligner), desarrollada por Obrusnik (2014). Hypal-4MUST está pensada para, de una manera relativamente intuitiva, combinar la alineación semiautomática de originales y traducciones, el etiquetado sintáctico y la anotación orientada a la traducción, de la que hablaremos en mayor detalle en la siguiente sección. Permite además la consulta del corpus, y posee dos entornos diferenciados, uno orientado a la investigación y otro destinado a la docencia (Granger y Lefer 2017).

A continuación se presenta una captura de pantalla de la interfaz de Hypal-4MUST (Figura 1) en la que puede verse cómo se alinean los textos originales con sus traducciones.

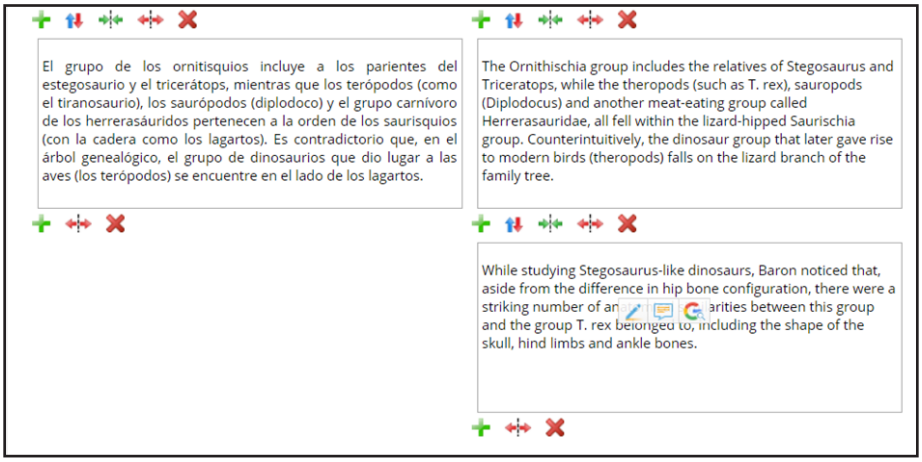

Figura 1. Interfaz de alineación en Hypal4MUST

3.1 Fases de recopilación del corpus en Hypal4MUST

El ciclo de recopilación de textos para el corpus utilizando la interfaz en línea Hypal4MUST consta de las siguientes fases.

En primer lugar se crea un corpus asociado al grupo de investigación, en nuestro caso, se denomina UCMA. A continuación se carga un texto original (en el caso de nuestro subgrupo la lengua de origen es siempre el inglés). Dicho texto se describe a través de una serie de metadatos estandarizados que se organizan en tres bloques, a saber:

a) datos sobre el texto de origen, información general (lengua, tipo de texto, autoría, número de palabras de la muestra,

b) palabras clave, y, por último,

c) información sobre la publicación (público meta, si existe una traducción de referencia…).

Una vez cargado el texto original y aprobado por las directoras del proyecto, se puede proceder a crear la tarea de traducción. Los participantes pueden entonces añadir sus propias traducciones a través de la interfaz de estudiantes, rellenando datos personales sobre su nivel de conocimientos lingüísticos y pericia traductora (que servirán para todas las traducciones de un mismo estudiante) y algunos datos específicos sobre el proceso de traducción (tiempo invertido,

herramientas documentales y de ayuda a la traducción empleadas, entre otras). Si los estudiantes no pudieran cargar sus traducciones, pueden hacerlo los propios profesores investigadores.

Cuando las traducciones están cargadas, se procede a su alineación semiautomática y posteriormente a la anotación de las traducciones por parte de los investigadores.

4. SISTEMA DE ANOTACIÓN ORIENTADA A LA TRADUCCIÓN (MUST-TAS)

Para la anotación de las traducciones se ha diseñado un sistema denominado TAS (*Translation-oriented Annotation System*), que en estos momentos se encuentra en fase de prueba. Este sistema se basa en las tipologías de anotación de errores en corpus de aprendices y en los estudios de traducción basados en corpus. Sin embargo, difiere de ellos en al menos dos aspectos. Por un lado, permite marcar tanto errores como aciertos. Por otro, da la posibilidad de señalar estrategias de traducción como la transposición, la simplificación o la explicitación, lo que posibilita por tanto investigaciones de corte más teórico (Granger y Lefer, 2017).

Para crear el sistema de anotación MUST-TAS, se ha tomado como ejemplo la taxonomía de errores CELTraC (Javora 2015 y Fictumova et al. 2017), que a su vez se basa en la taxonomía MeLLANGE (Kübler 2008). Así mismo, en la parte lingüística del sistema de anotación se han seguido las taxonomías de errores ICLE y FRIDA, diseñadas en el marco del Corpus Internacional ICLE de aprendices de inglés y el proyecto *French Interlanguage Database* (FRIDA) (Dagneaux et al., 1998; Granger, 2003).

El sistema de anotación TAS se compone de tres partes. En primer lugar, se anotan aspectos relacionados con la transferencia entre el texto original y el de llegada, lo cual facilita el marcado de las discrepancias entre el TO y el TT, o bien entre el TT y las directrices del encargo de traducción.

En segundo lugar, MUST-TAS contempla aspectos relacionados con la producción lingüística en el texto de llegada, es decir, aquellos que, con independencia de lo que diga el texto original, no son correctos desde el punto de vista lingüístico en el texto de llegada.

Por último, el sistema está diseñado para poder anotar los procedimientos utilizados para resolver problemas de traducción, que pueden observarse al

comparar el texto de origen y el texto meta. Esta tercera parte será especialmente útil para los investigadores y podrá utilizarse como herramienta didáctica para la traducción.

En estos momentos el sistema de anotación se encuentra todavía en fase de pruebas, pero sin duda será esta anotación de la traducción la que permita configurar un corpus con un alto valor como herramienta de investigación y didáctica.

5. CONTRIBUCIÓN DEL EQUIPO UCMA AL PROYECTO MUST

Nuestro equipo de investigación está compuesto por seis investigadores que imparten docencia en traducción en distintas universidades y en diferentes niveles educativos (grado, postgrado y preuniversitario). Todos trabajamos fundamentalmente en la combinación inglés-español, si bien puntualmente lo hacemos también con la combinación español-inglés en contextos de enseñanza del inglés y de la traducción.

A continuación se presentan los tipos de textos que recopilamos:

- Traducción audiovisual para el doblaje, la subtitulación para oyentes y para sordos.
- Textos humanísticos .
- Textos literarios.
- Textos de instituciones internacionales.
- Textos jurídicos y económicos.
- Textos generales.
- Textos científico-técnicos.
- Textos sobre diseño de modas.

6. EJEMPLOS DE ANÁLISIS DE TRADUCCIONES COMPILADAS POR EL EQUIPO DE LA UCM

La alineación de los textos mediante la interfaz Hypal4MUST permite cotejar los textos frase por frase en las dos lenguas de trabajo en cuestión. Nos vamos a centrar en uno de los textos jurídicos, concretamente un contrato de arrendamiento, cuya traducción puede plantear dificultades que abarcan aspectos tan diversos como la terminología, la fraseología, el formato, la ortotipografía, etc. En este apartado comentamos algunos de los errores más frecuentes cometidos por los estudiantes de un máster de traducción especializada que, en algunos casos, ya han cursado estudios de traducción y, por lo tanto, están familiarizados con algunas de las técnicas más utilizadas.

Vamos a mostrar algunos de los errores detectados en las alineaciones sobre la base de las categorías de errores indicadas en la clasificación descrita por el sistema TAS. Entre los errores más frecuentes nos centraremos en los que afectan al contenido, elemento léxico, gramática y ortotipografía.

El primer ejemplo muestra un error habitual que consiste en reproducir el formato de la lengua original en la lengua de destino. El fragmento introductorio ***THIS LEASE AGREEMENT*** figura en mayúsculas y en negrita en el original. Aunque en la traducción al español no es necesario calcar el formato, es frecuente verlo así: ***EL PRESENTE CONTRATO***.

El término *Whereas* se traduce normalmente por *considerando* en los textos jurídico-administrativos del ámbito internacional, pero en los contratos suele traducirse por la conjunción *que*, que se ajusta más a la forma y sentido de los contratos en español.

El tercer ejemplo concentra varios aspectos que suelen conducir a errores de traducción. Los verbos *made and entered* conforman un binomio fraseológico que se traduce erróneamente por *hecho y concertado*. Al tratarse de un binomio fraseológico, es decir, una expresión que algunos juristas consideran redundante porque usa dos términos con significado similar, se puede reducir a un solo término. La traducción literal de *made* por *hecho* no es la más acertada porque en español jurídico los contratos no se "hacen" sino que se "formalizan" o "celebran". Otro aspecto estaría relacionado con el tiempo verbal utilizado; en inglés el encabezamiento de los contratos se redacta en pasado o participio pasado, pero en español es en presente.

Estos son algunos de los errores más habituales en las primeras frases alineadas de la traducción cotejada. Una exposición sistematizada de este tipo de

errores les daría una mayor visibilidad y facilitaría el aprendizaje de los estudiantes, lo que contribuiría a su reducción.

7. RECAPITULACIÓN

En la presente contribución hemos presentado el proyecto MUST para la recopilación y anotación de un corpus de traducciones de estudiantes. Si bien el proyecto se encuentra todavía en una etapa preliminar y algunos de los procesos están todavía en fase de perfeccionamiento, no cabe duda de que el corpus, una vez compilado, alineado y anotado, será una fuente de referencia para los estudios de traducción basados en corpus, con implicaciones tanto para la didáctica de la traducción como para la investigación.

El grupo de la Universidad Complutense participa activamente en la compilación de este corpus con el fin de contribuir a su mejora, explotación y desarrollo.

REFERENCIAS BIBLIOGRÁFICAS

Dagneaux, E., Denness, S. y Granger S. 1998. "Computer-aided Error Analysis". *System: An International Journal of Educational Technology and Applied Linguistics,* 26: 163-174.

Fictumová, J., Obrusnik, A. y Štěpánková, K. 2017. "Teaching Specialized Translation: Error-tagged Translation Learner Corpora". *Sendebar*, 28: 209-241.

Granger, S. 2003. "Error-tagged learner corpora and CALL: a promising synergy". *CALICO*, 20: 465-480.

Granger, S. y Lefer, M.A. 2017. "Bridging the gap between learner corpus research and translation studies: The Multilingual Student Translation corpus". *4th Learner Corpus Research Conference*, Bolzano, Italia, 4-7 de octubre de 2017.

Jarova, S. 2015. *Defining an Error Typology: the Case of CELTraC*. Bachelor's Diploma Thesis. Masaryk: Masaryk University.

Johansson, S. 2007. *Seeing through Multilingual Corpora. On the Use of Corpora in Contrastive Studies*. Amsterdam: John Benjamins.

Kübler, N. 2008. "A comparable learner translator corpus: Creation and use". *LREC 2008 Workshop on Comparable Corpora*, 73-78.

Obrusnik, A. 2014. "Hypal: A user-friendly tool for automatic parallel text alignment and error tagging". *Eleventh International Conference Teaching and Language Corpora*. Lancaster, 20-23 de julio de 2014, 67-69.

Rica, J.P. En prensa. *Corpus-Based Studies and Audiovisual Translation: Subtitling*. Series: New Trends in Translation Studies (ed. J. Díaz Cintas). Frankfurt: Peter Lang.

Rica, J.P. y Braga, J. 2015. *Herramientas y técnicas para la traducción inglés-español: los textos literarios*. Madrid: Escolar y Mayo.

Rica, J.P., Albarrán, R. y García, B. 2014. "New approaches to audiovisual translation: the usefulness of corpus-based studies for the teaching of dubbing and subtitling". En E. Bárcena, T. Reads y J. Arus (eds.), *Languages for Specific Purposes in the Digital Era.* Berlin: Springer-Verlag, 303-322.