Propuesta recibida en octubre 2020 y aceptada para publicación en mayo 2021.

Corpus Paralelo de Español, Inglés y Chino y Análisis contrastivo del tiempo pasado del español a partir de corpus

Parallel corpus of Spanish, English and Chinese and corpus-based contrastive analysis of the past tense in Spanish

Hui-Chuan Lu
National Cheng Kung University

An Chung Cheng
University of Toledo

Meng-Hsin Yeh
National Cheng Kung University

Chao-Yi Lu
National Cheng Kung University

Ruth Alegre Di Lascio
National Cheng Kung University

Resumen

El presente estudio se dedica al desarrollo de un corpus paralelo trilingüe denominado CPEIC (Corpus Paralelo de Español, Inglés y Chino) cuyo fin es el de aportar conocimientos a las investigaciones sobre la traducción, el análisis contrastivo, el aprendizaje y la enseñanza de una lengua extranjera. Dicho CPEIC abarca las tres lenguas más habladas del mundo (español, inglés y chino) y contiene aproximadamente 4 millones de palabras. Basándose en el corpus paralelo desarrollado, se realizó un análisis contrastivo del tiempo pasado, el cual se expresa de manera diferente en las tres lenguas mencionadas. Los resultados obtenidos (a) avalan estudios previos sobre la relación entre el pretérito del español con el marcador aspectual chino "le", así como también la relación entre el imperfecto del español con "would" y "was/were+Ving" del inglés, (b) contradicen las presunciones con respecto a la conexión entre el imperfecto del español y el marcador aspectual chino "zhe", y (c) proporcionan una nueva perspectiva sobre la relación entre el pretérito del español y la voz pasiva en los tres idiomas.

Palabras clave

corpus paralelo, traducción, análisis contrastivo, tiempo pasado, aspecto, estudio a partir de corpus

Abstract

This study constructed a trilingual parallel corpus, the Parallel Corpus of Spanish, English and Chinese (CPEIC in Spanish), and used it to benefit research in translation, contrastive analysis, language learning, and teaching. The CPEIC contains the world's 3 most-spoken languages and comprises approximately 4 million words. Based on the construction result, a parallel corpus-based contrastive analysis was dedicated to the study of the past tense, which functions differently in the 3 languages. The results (a) support previous studies in associating the Spanish preterit and the Chinese aspectual marker "le", and in relating the Spanish imperfect with the English "would" and "was/were+Ving," (b) contradict assumptions for connecting the Spanish imperfect with the Chinese aspectual marker "zhe", and (c) offer new insights in uniting the Spanish preterit with the passive voice in all 3 languages.

Keywords

parallel corpus, translation, contrastive analysis, past tense, aspect, corpus-based

1. Introducción

La construcción del corpus y su respectiva investigación han jugado un papel crucial en el desarrollo de la lingüística durante los últimos 20 años. De acuerdo con los corpus enumerados en *Bookmarks for Corpus-based Linguistics* (Lee, 2010), la mayoría son corpus en inglés y fueron construidos con un propósito específico. Los corpus paralelos relacionados con lenguas distintas al inglés han atraído escasa atención. Esto puede ser atribuido a las dificultades de obtención y manejo de datos paralelos provenientes de varios idiomas. Sin embargo, debido a que la creación de un corpus paralelo involucra múltiples lenguas, puede proporcionar diversos valores académicos y de aplicación que no se encuentran en otros tipos de corpus. En comparación con los corpus monolingües o comparables, un corpus paralelo multilingüe puede ayudar al estudio de la traducción y al análisis contrastivo, con el fin de controlar las variables ocultas en el contenido no paralelo, y así llegar a una conclusión más objetiva y convincente.

DOI: 10.21814/lm.13.1.340

A diferencia del uso de las concordancias paralelas ya existentes o las herramientas del corpus, un corpus paralelo con funciones de búsqueda puede obtener resultados avanzados de manera más efectiva, como, por ejemplo, datos etiquetados y alineamiento de palabras, lo cual reduciendo así el tiempo de búsqueda. Entre los corpus paralelos existentes (Lee, 2010), no existe ningún corpus paralelo simultáneo entre los tres idiomas más hablados del mundo: inglés, español y chino, aunque cabe destacar que existen alineados de dos en dos (inglés–español, inglés–chino). Por lo tanto, la creación de un corpus paralelo trilingüe de este tipo facilitaría la investigación sobre la lingüística contrastiva y podría ser aplicada al estudio de la adquisición de una segunda lengua o el multilingüismo.

Basándose en el corpus paralelo trilingüe creado en este estudio, se examinó una característica lingüística en particular: el tiempo pasado. Este tiempo verbal se emplea de manera diferente en estos tres idiomas. El tiempo pasado en el español posee dos posibles conjugaciones para el verbo: el pretérito y el imperfecto, mientras que en el inglés sólo se cuenta con una conjugación para el verbo: el tiempo pasado ("past tense"). El chino, por su parte, no posee inflexiones verbales. Esta característica lingüística fue investigada debido a que el tiempo pasado en el español resulta ser una de las reglas gramaticales más difíciles de aprender para los estudiantes taiwaneses, cuya lengua materna (L1) es el chino, y cuya segunda lengua (L2) y primera lengua extranjera es el inglés. Por lo tanto, el corpus paralelo trilingüe construido refleja el contexto bajo el cual el idioma español es asimilado en Taiwán. La creación del corpus paralelo trilingüe, los hallazgos del estudio de la traducción y el análisis contrastivo utilizando el corpus construido, proporcionan pautas útiles para la traducción del español y la enseñanza o el aprendizaje de otros idiomas.

Este artículo está organizado de la siguiente manera: en la segunda sección, se revisan estudios previos y concurrentes relacionados al tema. En la tercera sección, se presenta la creación y evaluación del corpus paralelo trilingüe. En la cuarta sección, se provee un estudio de traducción y análisis contrastivo basado en el corpus paralelo construido y en sus implicaciones pedagógicas. Finalmente, en la quinta sección se brinda una conclusión del artículo.

2. Revisión literaria

2.1. Corpus Paralelo

Utilizando el término "corpus paralelo" como filtro para la búsqueda en la base de datos de *Linguistics and Language Behavior Abstracts* (LLBA) se observó el aumento de publicaciones de investigaciones relacionadas a dicho tema en los últimos 40 años. La gran mayoría de estos estudios se han enfocado en la construcción de un corpus paralelo. Por lo general, estos estudios se han centrado en lenguas indoeuropeas. En consecuencia, las combinaciones de lenguas asiáticas son relativamente escasas, encontrándose sólo la del japonés–chino (Ma et al., 2004).

En el proceso de creación del corpus paralelo, la alineación entre los diferentes idiomas resultó ser el problema más difícil de superar, igual que lo previamente mencionado por Sun et al. (2002). En la actualidad, la alineación de oraciones en el proceso de creación de un corpus paralelo no presenta el mismo nivel de complejidad que la alineación de palabras, cuyo proceso require de más pasos, pero que, a su vez, obtiene mejores resultados de correspondencia semántica. Con el fin de obtener un conocimiento general sobre el desarrollo actual de los corpus paralelos, se evaluaron aquellos relacionados a los tres idiomas más hablados del mundo: español, inglés y chino. Una vez evaluados los resultados, dos tipos de corpus paralelos fueron identificados. El primer tipo compila datos en paralelo en una carpeta, permitiendo la descarga de dichos datos; no obstante, nos enfocamos en el segundo tipo, el cual proporciona funciones de búsqueda y otras funciones más avanzadas.

La primera categoría incluye (1a) the European Parliament Proceedings Parallel Corpus, o, Corpus Paralelo de las Actas del Parlamento Europeo (Koehn, 2005) y (1b) Multilingual and Parallel Corpora European Commission, o, la Comisión Europea de Corpus Paralelos y Multilingües (MLCC por sus siglas en inglés) (ELRA, 1996). El primer corpus, (1a), contiene fuentes procedentes de las actas del Parlamento Europeo (1996–2009), incluyendo corpus paralelos del inglés (49.093.806 palabras) y español (51.575.748 palabras), con 1.965.734 oraciones paralelas en total. El propósito para la creación de dicho corpus fue el de desarrollar un sistema de traducción automática, el cual no dispone de ninguna función de búsqueda. Los datos del segundo corpus, (1b), se obtuvieron del Diario Oficial de las Comunidades Europeas (6.000.000–9.000.000 palabras para cada idioma), que cobra una tarifa por su uso. Los siguientes corpus cuentan

con interfaces de usuario y funciones de búsqueda de palabras: (2a) Corpus Paralelo de Análisis Contrastivo y Traducción Inglés–Español, P-ACTRES (Rabadán, 2002), que consta de libros, editoriales de periódicos, artículos de revistas y ensayos (2.500.000 palabras). El propósito para la creación de este corpus fue el de realizar un análisis contrastivo del inglés y el español, las aplicaciones del inglés como lengua extranjera y la enseñanza de la traducción. Las búsquedas de palabras y POS están disponibles. A diferencia de otros corpus, (2b) Open Source Parallel Corpus, OPUS (Tiedemann, 2012) ofrece una búsqueda pública de datos de traducción entre el español (400.000–500.000 palabras) y el inglés (500.000 palabras). Como resultado de dicha búsqueda se presenta la alineación de las oraciones provenientes de estos dos idiomas. Por su parte, (2c) English–Chinese Parallel Concordance, E–C Conco d (Lixun, 2001) contiene 1.878.795 palabras en inglés y 3.152.866 caracteres chinos, derivados de novelas, documentos legales, artículos académicos, cuentos de hadas, discursos, ensayos, y fábulas. El mismo dispone de funciones de búsqueda de palabras y como resultado se presenta la alineación de oraciones entre estos dos idiomas. (2d) El Corpus Paralelo Inglés–Chino de Babel, ParaConc (McEnery & Xiao, 2005) comprende 253.633 palabras en inglés y 287.462 caracteres chinos, y cuyas fuentes de datos son *World of English* y *Time*. Utilizando ParaConc, los resultados de búsqueda generados son oraciones alineadas etiquetadas con POS.

Algunos corpus paralelos ya existentes se encuentran en etapa de recopilación de información; sin embargo, ninguno trabaja simultáneamente con los tres idiomas, tampoco presentan una interfaz de usuario, ni consideran las búsquedas de palabras o las de POS. Esta notable falta de un corpus paralelo español–inglés–chino motivó la realización del presente estudio. Con base en la revisión de estudios previos (Corpas Pastor, 2003; Castillo Rodríguez, 2009; Baker, 1995; Malmkjær, 2005; Rabadán et al., 2009; Dimitrova et al., 2010), se construyó un corpus paralelo con el objetivo de realizar un análisis de contraste enfocado en las características lingüísticas del tiempo pasado del español, lo cual beneficiará a investigaciones futuras en el área del aprendizaje de idiomas.

2.2. Tiempo Pasado

Mediante un análisis de contraste basado en los corpus paralelos, se investigaron las conjugaciones de los tiempos pasados del español y sus interacciones y diferencias con las del inglés y del chino. Dichas diferencias se pueden observar en las siguientes tres oraciones:

(1) Cuando conversábamos, sonó el teléfono.
(2) When we were talking, the telephone rang.
(3) 當 我們正在聊天時， 電話 響了。
Dang women zhengzai liaotianshi dianhua xiangle

En el español, el tiempo pasado está marcado rotundamente; la morfología flexiva indica tanto el tiempo como el aspecto del verbo. El pretérito codifica la perfectividad, mientras que el imperfecto codifica la imperfectividad. Como se muestra en la oración (1), la terminación flexiva "-ábamos" es utilizada como marcador del imperfecto, y "-ó" es un marcador que simboliza el pretérito. En el inglés, las distinciones aspectuales de los verbos son menos notorias que las distinciones temporales utilizadas por los hablantes nativos del inglés. En el inglés, el contraste aspectual más evidente se encuentra entre el progresivo y el perfectivo al momento de utilizar el pasado progresivo y el pasado simple. En la oración (2), la terminación flexiva "-ing" indica la acción progresiva y "rang" indica el pasado simple. Por otro lado, en el chino no se marcan contrastes en tiempos y aspectos verbales a través de la morfología. Los hablantes nativos del chino son capaces de identificar y comprender los distintos tiempos verbales a través de adverbios, adjuntos, argumentos y referencias contextuales. En la oración (3), "zai" es un marcador aspectual que denota un significado progresivo y "le" es un marcador de aspecto perfectivo.

Con base en la literatura previa y la observación gramatical, las variables que juegan un papel en la interacción de los tiempos verbales entre estos tres idiomas incluyen: los aspectos gramaticales del español (el aspecto del pretérito y el marcador imperfecto), los objetos de los verbos, la negación, los adverbios y conjunciones temporales, los tiempos verbales del inglés (el pasado simple y progresivo) y los marcadores aspectuales del chino ("guo, zai, zhe, le") (Bardovi-Harlig, 2000; Robison, 1990; Salaberry, 2002, 2011; Vendler, 1967; Xiao & McEnery, 2004). Estas variables fueron examinadas con un énfasis especial en el corpus.

3. Metodología

3.1. Creación del CPEIC

Para diferenciarse y resaltar entre los corpus paralelos ya existentes, se establecieron los siguientes objetivos relacionados al idioma, la alineación, el etiquetado, la función de búsqueda y la presentación de resultados: (a) el corpus paralelo trilingüe de español–inglés–chino, (b) el etiquetado POS y la alineación de palabras y (c) la búsqueda simultánea en varios idiomas y a través de palabras clave. El procedimiento de creación consta de cuatro pasos principales: (a) la compilación de datos, (b) el etiquetado POS, (c) la alineación de palabras y (d) la programación de software, como se muestra en la Figura 1.

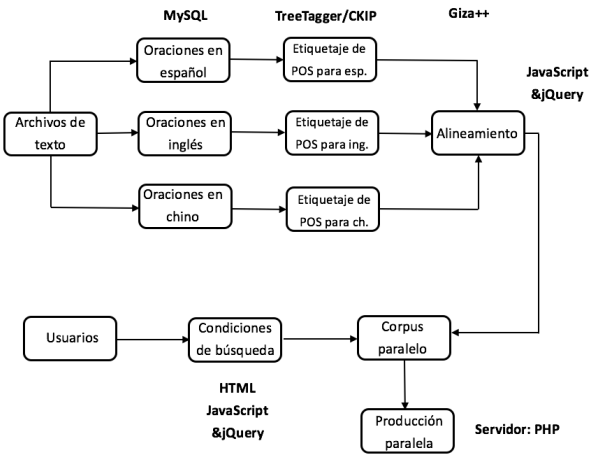

Figura 1: Diagrama de flujo de creación de CPEIC.

Se comenzaron a recopilar los datos paralelos para la CPEIC en el 2007 a partir de tres fuentes: (a) la Biblia, (b) cuentos de hadas y (c) formas escritas y habladas de documentos de las Naciones Unidas (ONU) en los tres idiomas (español, inglés y chino) con el objetivo de incluir un amplio repertorio de temas y de tipos de texto. Las fuentes encontradas fueron escasas debido a que existen pocos textos paralelos en español y chino que sean gratuitos y de fácil disposición en Internet. El manejo de datos varió según la fuente debido a que los textos que provienen de distintos sitios web cuentan con diversos formatos de sistema en Internet. Hasta la fecha, el CPEIC contiene 1.193.418 palabras en español, 1.199.715 palabras en inglés y 1.200.914 caracteres chinos.

En las etapas iniciales, los datos recopilados para el español e inglés se importaron a MySQL para etiquetarlos en POS utilizando TreeTagger (Solorio & Liu, 2008), mientras tanto, para el chino se utilizó el Procesamiento de Conocimiento e Información en Chino (CKIP), sistema de segmentación de palabras en chino (Ma & Chen, 2003). Posteriormente, las palabras se alinearon del idioma A al B, y seguidamente del idioma B al A utilizando Giza++ (Och & Ney, 2000). Este paquete de software procesa abundantes datos con el fin de minimizar los errores al alinear la lengua fuente y la lengua meta; y se basa en el modelo de probabilidad optimizado para predecir las correspondencias semánticas más posibles. Se diseñó una interfaz de usuario de entorno web utilizando JavaScript y jQuery, mientras que el servidor fue programado con el Preprocesador de Hipertexto (PHP). La interfaz de usuario se muestra en la Figura 2.

Figura 2: Interfaz de usuario CPEIC.

3.2. Evaluación del CPEIC

Tras culminar la primera etapa de la construcción del CPEIC, los resultados obtenidos fueron evaluados, como se indica en esta sección. De acuerdo con la evaluación realizada, el uso del corpus paralelo construido para la búsqueda de datos contrastivos puede ser siete veces más rápido que el tradicional. Al examinar los objetivos establecidos y comparar el corpus con los ya existentes, el CPEIC creado se destaca por las siguientes razones: (a) es un corpus paralelo que resuelve el problema de compatibilidad de tres idiomas, (b) los datos están etiquetados con POS y alineados con palabras para la función de búsqueda y presentación de resultados, y (c) la búsqueda en la interfaz de usuario puede ser simultáneamente trilingüe e incluir múltiples palabras clave y consultas compuestas con una velocidad de búsqueda mejorada y una frecuencia de cierre reducida.

Además, el corpus construido fue evaluado calculando el porcentaje de precisión del etiquetado POS de cada idioma y la alineación de palabras entre dos idiomas. Los resultados indicaron que las tasas de precisión del etiquetado POS para la Biblia, los cuentos de hadas y los documentos de la ONU escritos y orales excedieron el 93 % para inglés y español, mientras que variaron del 76 % al 93 % para el chino en las diversas

fuentes de datos. Esta diferencia entre la tasa de precisión de etiquetado del español e inglés con la del chino podría ser el resultado de los diferentes sistemas de etiquetado utilizados para las fuentes de datos: TreeTagger para inglés y español, y CKIP para chino. Si bien es cierto que este corpus paralelo trilingüe puede simplificar el proceso de búsqueda en una investigación proporcionando abundantes datos, múltiples funciones y un motor de búsqueda rápido, los sistemas adoptados mostraron errores en los resultados de búsqueda. En la etapa actual, los resultados parcialmente incorrectos en el etiquetado POS y la alineación de palabras han requerido una verificación posterior y corrección manual a fin de obtener mejores resultados.

Finalmente, evaluamos el valor del autoaprendizaje desde la perspectiva de los usuarios, identificando las ventajas de la aplicación del corpus construido. Entre estas ventajas se encuentra el poder diferenciar dos elementos similares al proporcionar niveles de oración y párrafo para que así los alumnos obtengan más detalles y entendimiento léxico. En cuanto a la satisfacción del usuario, según la experiencia de búsqueda, más del 69 % de los participantes informaron que la CPEIC fue útil para ayudarlos a adquirir conocimientos lingüísticos, en particular para proporcionar traducción paralela a través de similitudes o diferencias de estructuras y expresiones léxicas.

4. Estudio contrastivo del tiempo pasado basado en el corpus paralelo

4.1. Objetivo y preguntas de investigación

El estudio de la traducción y el análisis contrastivo mejoran nuestro conocimiento de la gramática universal y los parámetros específicos para cada idioma, lo cual también aumenta nuestra comprensión de la transferencia de estos. Debido a que L1 (lengua materna), L2 (primera lengua extranjera) y L3 (segunda lengua extranjera) son tres idiomas paralelos para los estudiantes taiwaneses de español, el estudio de la traducción y el análisis contrastivo de estos idiomas ayuda a examinar la influencia de la primera y segunda lengua en la adquisición de la tercera lengua. Por lo tanto, este estudio abordó la siguiente pregunta de investigación: Basado en el estudio de la traducción y el análisis contrastivo de un corpus paralelo trilingüe, ¿cuáles son las variables lingüísticas que podrían ser asociadas con los dos aspectos gramaticales del tiempo pasado en español?

4.2. Metodología

Al analizar los elementos léxicos y las estructuras sintácticas de los resultados alineados, las similitudes y diferencias se compararon y contrastaron observando los tres lenguajes paralelos y la interacción entre ellos. Con el objetivo de determinar los factores relevantes que podrían afectar la selección de dos aspectos gramaticales diferentes en el tiempo pasado del español, las posibles variables incluidas en el examen fueron las siguientes: los adverbios y las conjunciones temporales en español, las negaciones, los objetos de los verbos y la voz pasiva; el tiempo pasado en inglés (tiempo pasado simple, progresivo, y futuro en tiempo pasado) y voz pasiva; y cuatro marcadores de aspecto en chino ("guo, zai, zhe y le").

Los datos analizados se extrajeron del corpus trilingüe paralelo construido, (CPEIC), que incluía las tres fuentes antes mencionadas: la Biblia, los cuentos de hadas y las formas escritas y orales de los documentos de la ONU. En este estudio se examinaron un total de 198.386 palabras y se analizaron 2.160 verbos en tiempo pasado en español y los elementos correspondientes en inglés y chino.

4.3. Resultados y debate

Mediante la aplicación del análisis contrastivo de los datos paralelos provenientes de diversas fuentes, se obtuvieron los siguientes resultados. El resultado más relevante demostró la contradicción entre los marcadores de aspecto de acción progresiva en chino "zhe" con el imperfecto del español. El marcador de aspecto "zhe" se empleó de manera diferente a lo que se esperaba. La mayor correspondencia al marcador progresivo "zhe" del chino fue alcanzada por el pretérito del español y no por el imperfecto, en el caso de la biblia el pretérito se utilizó en un 80 % de los casos y en los cuentos de hada en un 59 %.

Por lo tanto, seguir la suposición convencional de que el marcador progresivo chino "zhe" está asociado con la forma imperfecta del español crearía una correspondencia errónea entre el español y el chino. Sin un estudio paralelo basado en un corpus, esta evidencia contraria no se habría descubierto. Por lo tanto, los hallazgos del análisis contrastivo translingüístico verifican el conocimiento gramatical previo.

Al continuar con el estudio paralelo se realizó otro descubrimiento destacado que estudios previos no han discutido. Entre los factores que inciden en la selección entre los dos aspectos del tiempo pasado del español, la voz pasiva en los tres idiomas (*ser*+PP(100 %) en español,

be+PP(100 %) en inglés, y *bei* (más del 67 %) en el chino) está altamente asociada al uso del pretérito en el español y no al imperfecto.

Además, el resultado proporciona evidencia basada en el corpus para respaldar el conocimiento gramatical previamente establecido. Entre las variables examinadas en inglés, el tiempo progresivo, "*be* + Ving" (más del 93 %) y el futuro en tiempo pasado "*would* + V" (más del 92 %) tienen una fuerte relación con el tiempo imperfecto español. En cuanto al chino, el marcador de aspecto perfectivo "le" tiene una alta tendencia (más del 82 %) a aparecer con el pretérito español.

También derivamos ciertos principios concluyentes para la enseñanza y el aprendizaje del español. En primer lugar, el pretérito en español (87 %, 60 %, 89 % y 88 % para la Biblia, los cuentos de hadas y las formas escritas y orales de los documentos de la ONU, respectivamente) generalmente aparece con más frecuencia que la forma imperfecta, lo que sugiere una mayor frecuencia del pretérito en el input del aprendizaje de este idioma. Además, Lu et al. (2015) propusieron que los estudiantes de segunda lengua (estudiantes de habla inglesa y mandarín) adquieran el pretérito español antes que el imperfecto. Por lo tanto, proponemos enseñar el tiempo pasado en español en función del orden de frecuencia de cada aspecto y la adquisición del alumno de dicho aspecto. Los estudiantes deben aprender los verbos de alta frecuencia de acuerdo con el siguiente orden de aspecto gramatical: (a) verbos utilizados exclusivamente en pretérito, (b) verbos utilizados exclusivamente en el imperfecto, (c) verbos utilizados con más frecuencia en pretérito que en el imperfecto, y (d) verbos utilizados con más frecuencia en el imperfecto que el pretérito. Además, ciertas palabras clave están relacionadas con aspectos específicos, como por ejemplo, la aparición de la conjunción temporal "mientras" y el adverbio temporal "siempre" garantizan el uso del aspecto imperfecto (100 %) en los textos de cuentos de hadas, mientras que la mayoría de las conjunciones temporales españolas, "cuando" o "when" (en inglés), se asocian con la forma pretérita en la Biblia (88 %), los cuentos de hadas (76 %) y en la forma oral de los documentos de la ONU (91 %).

Finalmente, además de las generalizaciones concluyentes antes mencionadas, la relación entre los factores y las fuentes del texto a veces presentan cambios. Por ejemplo, las negaciones *no* ("no") y *ni* ("neither") tienden a aparecer más frecuentemente con el imperfecto español en los cuentos de hadas y en los escritos de los documentos de la ONU, mientras que muestran comportamientos diferentes en la Biblia y en los documentos orales de la ONU. Por lo tanto, la selección de materiales didácticos para propósitos específicos debe tener en cuenta géneros y formas.

4.4. Enseñanza y aprendizaje

Un corpus paralelo trilingüe puede ser utilizado en investigaciones lingüísticas y presentar numerosas implicaciones pedagógicas para la traducción, enseñanza y aprendizaje de una segunda lengua para los estudiantes de estas lenguas. Con base en la traducción y los resultados contrastivos a través de búsquedas utilizando el corpus trilingüe construido, los profesores pueden diseñar materiales didácticos y llamar la atención de los estudiantes de español L3 para fortalecer los efectos positivos de L1 (chino) y L2 (inglés) y evitar sus efectos negativos según las similitudes y las diferencias entre las tres lenguas.

Para los principiantes, en comparación con los corpus monolingües, el presente corpus multilingüe se puede utilizar para buscar ejemplos o correlaciones que correspondan con la traducción de la lengua materna o la primera lengua extranjera para facilitar la determinación de significados (por ejemplo, *ser* o "be" vs. *estar* o "be"). Para estudiantes más avanzados, el corpus construido puede ayudar en la aclaración de las diferencias triviales entre las formas relacionadas mediante la búsqueda transversal de dos aspectos (por ejemplo, *fueron* vs. *eran* o *estaban* vs. *estuvieron*). Además, dado que, el CPEIC puede obtener de manera sistemática y eficiente patrones, características y categorías clasificadas basándose en la frecuencia del uso del pretérito y del imperfecto, sugerimos que los profesores consideren la frecuencia de aparición de los verbos en los datos del idioma nativo para diseñar sus materiales de enseñanza con verbos de alta frecuencia (*decir, salir, estar* "say, leave, be") siguiendo la secuencia de aprendizaje (sólo verbos en pretérito > sólo verbos en imperfecto > más pretérito que imperfecto > más imperfecto que pretérito) y obteniendo una lista de palabras (*soler* "usually do" en la de únicamente imperfecto; *ser, tener* "be, have" en la lista del pretérito y del imperfecto) con ejemplos paralelos auténticos.

Con respecto a las fuentes del texto, debido a que varios tipos de literatura y temas están asociados con diversas selecciones de un aspecto verbal, al elegir materiales para los estudiantes, los profesores deben considerar las características y la tendencia de uso dentro de los géneros y temas en los materiales elegidos. Los verbos del

español en tiempo pasado en los cuentos de hadas ofrecen una plataforma adecuada para comparar dos aspectos gramaticales diferentes porque ambas formas aparecen en una base cuasi frecuente (60 % en forma pretérita frente a 40 % en forma imperfecta). Además, los documentos de la ONU que ofrecen tanto textos escritos como orales permiten a los alumnos comparar y contrastar las similitudes y diferencias entre los dos formatos. Por último, aunque el lenguaje utilizado en los textos bíblicos se considera antiguo y el vocabulario y la gramática pueden no ser óptimos para facilitar el aprendizaje de los estudiantes, los temas religiosos pueden atraer la atención de aquellos que estén interesados en la religión.

5. Conclusión

En resumen, en cuanto a la creación del CPEIC, hemos completado la primera etapa en la construcción de un corpus trilingüe. La cantidad total de datos es de aproximadamente 4 millones de palabras. Dicho corpus difiere de los corpus paralelos existentes y se caracteriza por la información etiquetada en POS y la alineación de palabras de los tres idiomas más hablados en todo el mundo: español, inglés y chino. El corpus proporciona una plataforma para estudios académicos y se puede utilizar como una herramienta para ayudar a los instructores en la enseñanza y para facilitar a los estudiantes multilingües el proceso de aprendizaje en la conexión de forma y significado.

Mediante el análisis contrastivo y la traducción del tiempo pasado, el CPEIC obtuvo de manera sistemática y eficiente patrones, rasgos y categorías clasificadas de dos aspectos gramaticales del español, el pretérito y el imperfecto, según la frecuencia de uso. Con base en el resultado del análisis, que incluye evidencia tanto respaldada por estudios anteriores como contradictoria a ellos, la secuencia de aprendizaje y las sugerencias para la selección de textos proporcionan referencias para los maestros y estudiantes de español L3 en Taiwán.

Reconocimiento

Extendemos nuestro agradecimiento al Ministerio de Ciencia y Tecnología de Taiwán por su generoso apoyo a la subvención de este proyecto (MOST 101-2410-H-006-088-MY2), el apoyo técnico del equipo de Ciencias de la Computación e Ingeniería de la Información (CSIE) en la Universidad Nacional Cheng Kung en Taiwán, y a los asistentes de investigación.

Referencias

Baker, Mona. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target. International Journal of Translation Studies* 7(2). 223–243. 10.1075/target.7.2.03bak.

Bardovi-Harlig, Kathleen. 2000. Tense and aspect in second language acquisition: Form, meaning, and use. *Language Learning: A Journal of Research in Language Studies* 50. 1.

Castillo Rodríguez, Cristina. 2009. La elaboración de un corpus ad hoc paralelo multilingüe. *Tradumàtica* (7). on–line.

Corpas Pastor, Gloria. 2003. TURICOR: compilación de un corpus de contratos turísticos (alemán, español, inglés, italiano) para la generación textual multilingüe y la traducción jurídica. En *Panorama actual de la investigación en traducción e interpretación*, 373–384.

Dimitrova, Ludmila, Violetta Koseska-Toszewa, Danuta Roszko & Roman Roszko. 2010. Application of multilingual corpus in contrastive studies (on the example of the Bulgarian-Polish-Lithuanian parallel corpus). *Cognitive Studies* (10). 10.11649/cs.2010.013.

ELRA, European Language Resources Association. 1996. Multilingual and Parallel Corpora. https://tinyurl.com/ycz4x9ky.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. En *10th Machine Translation Summit*, 79–86.

Lee, David. 2010. Bookmarks for corpus-based linguists. Available from World Wide Web: http://devoted.to/corpora.

Lixun, Wang. 2001. Exploring parallel concordancing in English and Chinese. *Language Learning & Technology* 5(3). 174–184.

Lu, Hui-Chuan, An Chung Cheng & Sheng-Yun Hung. 2015. La adquisición del tiempo-aspecto en L3 para los aprendices taiwaneses. *Círculo de Lingüística Aplicada a la Comunicación* 63. 200–217. 10.5209/rev_CLAC.2015.v63.50175.

Ma, Qing, Kyoko Kanzaki, Yujie Zhang, Masaki Murata & Hitoshi Isahara. 2004. Self-organizing semantic maps and its application to word alignment in Japanese–Chinese parallel corpora. *Neural networks* 17(8-9). 1241–1253. 10.1016/j.neunet.2004.07.011.

Ma, Wei-Yun & Keh-Jiann Chen. 2003. Introduction to CKIP Chinese word segmentation

system for the first international Chinese word segmentation bakeoff. En *Second SIGHAN workshop on Chinese language processing*, 168–171. 10.3115/1119250.1119276.

Malmkjær, Kirsten. 2005. *Linguistics and the language of translation*. Edinburgh university press.

McEnery, Tony & Richard Xiao. 2005. The babel English-Chinese parallel corpus. http://www.lancs.ac.uk/fass/projects/corpus/babel/babel.html.

Och, Franz Josef & Hermann Ney. 2000. Improved statistical alignment models. En *38th Annual Meeting on Association for Computational Linguistics*, 440–447. 10.3115/1075218.1075274.

Rabadán, Rosa, Belén Labrador & Noelia Ramón. 2009. Corpus-based contrastive analysis and translation universals: A tool for translation quality assessment English → Spanish. *Babel* 55(4). 303–328. 10.1075/babel.55.4.01rab.

Rabadán, Rosa. 2002. Análisis contrastivo y traducción inglés-español: el programa ACTRES, 35–55.

Robison, Richard E. 1990. The primacy of aspect: Aspectual marking in English interlanguage. *Studies in second language acquisition* 12(3). 315–330. 10.1017/S0272263100009190.

Salaberry, Maximo Rafael. 2002. Tense and aspect in the selection of Spanish past tense verbal morphology. *Language acquisition and language disorders* 27. 397–416. 10.1075/lald.27.16sal.

Salaberry, Maximo Rafael. 2011. Assessing the effect of lexical aspect and grounding on the acquisition of L2 Spanish past tense morphology among L1 English speakers. *Bilingualism: Language and Cognition* 14(2). 184–202. 10.1017/S1366728910000052.

Solorio, Thamar & Yang Liu. 2008. Part-of-speech tagging for English-Spanish code-switched text. En *Conference on Empirical Methods in Natural Language Processing*, 1051–1060.

Sun, Le, Song Xue, Weimin Qu, Xiaofeng Wang & Yufang Sun. 2002. Constructing of a large-scale Chinese-English parallel corpus. En *3rd workshop on Asian language resources and international standardization*, 1–8.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. En *Language Resources Evaluation Conference (LREC)*, 2214–2218.

Vendler, Z. 1967. Verbs and times text. *Linguistics in Philosophy. Ithaca, NY: Cornell University Press* 97.

Xiao, Richard & Tony McEnery. 2004. *Aspect in Mandarin Chinese: A corpus-based study*, vol. 73. John Benjamins Publishing.