外國語文研究第九期
2009 年 1 月 頁 135~156

學習者西英書面語平行語料庫之建構

盧慧娟*、呂羅雪**

中文摘要

本論文旨在初始建置一個以台灣西語學習者為語料收集對象，帶有註記的西英平行書面語語料庫。首先，我們將逐步擴增「台灣英西平行語料庫」之語料筆數及其多元性。此學習者語料庫所收集的語料為在台灣以英語和西語分別為第二語和第三語之中介語，以作為日後探究「中介語的發展及變異性」領域中相關議題的基礎，達到便利學習者多語中介語之分析研究的目的。同時，針對語料註記正誤與轉移變數的標記，以提升語料庫的利用價值。此外，透過現有語料庫工具之搭配與輔助，強化語料的檢索功能。最後，我們希望藉由語料庫建構及應用模式的呈現與經驗分享，推動語料庫語言學理論與應用在台灣外語教學領域的發展。

關鍵字：語料庫建構、註記、中介語、第三語、西班牙語學

*成功大學外國語文學系教授。
** 靜宜大學西班牙語文學副教授。
97.10.24 到稿　97.12.9 通過刊登

The Construction of Spanish-English Learners' Written Parallel Corpus (CPATEI)

Hui-Chuan Lu* & Hsueh Lo Lu**

Abstract

This paper aims to begin to construct an annotated corpus of Taiwanese learners' Spanish-English parallel written texts. Firstly, we have been gradually expanding the quantity and variety of parallel written data since 2006. This learners' corpus mainly compiles data of interlanguage from Taiwanese learners who learn English as second language and Spanish as third language. Meanwhile, the values of corpus have been enhanced by annotating the corrections of learners' errors and variables of L1/L2 transfer with the final goal of benefiting the analysis of related fields such as the study of multilingual interlanguage and those on the research of development and variability in interlenguage. In addition, the searching function can be improved through the concordance assisted by combining different existing corpus tools. Finally, we hope to promote the development of corpus linguistics theory and its practical application on foreign language teaching in Taiwan by sharing our experience in constructing corpora.

Keywords: corpus construction, annotation, interlanguage, L3, Spanish learning

* Professor, Department of Foreign Languages & Literature, National Cheng Kung University.
** Associate Professor, Department of Spanish Language and Literature, Providence University

外國語文研究第九期
2009 年 1 月 頁 135~156

Creación del corpus paralelo de aprendices taiwaneses de español e inglés (CPATEI)

Hui-Chuan Lu* & Hsueh Lo Lu**

Resumen

El principal objetivo de nuestro trabajo es la creación inicial de un corpus anotado y paralelo de los aprendices taiwaneses de español e inglés. El procedimiento que hemos seguido es el siguiente: En primer lugar, continuamos construyendo el CPATEI para ampliar tanto la cantidad como la variedad de los datos obtenidos. Este corpus se basa en la recopilación de datos paralelos de interlengua del inglés como L2 y del español como L3 sobre los aprendices taiwaneses, con el fin de proporcionar recursos útiles con los que se pueda profundizar en investigaciones relacionadas con el desarrollo y la variación de interlengua o con el área de estudio donde existe más de una interlengua. En segundo lugar, incluimos etiquetas, anotaciones y transformaciones lingüísticas para enriquecer el valor del corpus. En tercer lugar, mostramos la consulta de los resultados utilizando las herramientas existentes. Por último, esperamos que la creación del corpus y el proceso de su construcción en nuestra investigación pueda ser de utilidad tanto en la teoría como en la aplicación de la Lingüística de Corpus adoptada en la enseñanza de las lenguas extranjeras en Taiwán.

Palabras clave: creación de corpus, anotación, interlengua, L3, aprendizaje de español

* Profesora catedrática, Departamento de Lengua y Literatura Extranjeras, National Cheng Kung University.
** Profesora asociada, Departamento de Lengua y Literatura Españolas, Universidad Providence.

Estudio de un corpus paralelo de escritura del español: Creación de CPATEI[1]

Hui-Chuan Lu & Hsueh Lo Lu

1. Introducción

En la historia de la Lingüística de Corpus, los estudios, tanto en su vertiente teórica como aplicada, empezaron a desarrollarse hace cuarenta años y la cantidad se ha incrementado aún más en la última década. Esto demuestra que los estudios relacionados con los corpus ocupan un puesto importante en la corriente lingüística. La creación de corpus anotados desempeña un papel esencial en los estudios basados en corpus, en el sentido de que no sólo satisface la necesidad de ser capaz de ofrecer datos representativos, sino que también guía el desarrollo de otras áreas lingüísticas, tales como la adquisición de la segunda lengua. Sin embargo, en el campo de la lingüística española de corpus en Taiwán, los recursos no han sido fruto de una colaboración como consecuencia de la falta de herramientas y sistemas integrados de manera que siguen desarrollándose o utilizándose de forma individual. Por ejemplo, en las 4 universidades taiwanesas donde hay Departamento de Español, 28 profesores (21.4%)[2] se especializan en la lingüística teórica o aplicada. Estos profesores dedican su tiempo a recopilar datos de producción escrita (en general, las composiciones) de los estudiantes, para realizar investigaciones didácticas o lingüísticas. Resulta penoso que cada docente dedique su tiempo y energía a recopilar una cantidad de datos, sólo para una investigación o para el estudio de cierto tema específico. Después de terminar tal

[1] El presente trabajo originalmente se presentó en el XLIII Congreso Internacional de la AEPE (2008) que tuvo lugar en Madrid. Este estudio es parte del proyecto integrado piloto titulado “Desarrollo y variación de interlengua” patrocinado por el Centro de Investigación para Humanidades del Consejo Nacional de Ciencias de Taiwán.

[2] El resultado se calcula mediante la consideración de los trabajos escritos publicados por los profesores de la especialización considerada.

investigación, los datos se dejan sin reutilizarse ni aprovecharse otra vez. Como consecuencia de esto, la cooperación entre las distintas investigaciones académicas se elimina tanto por la escasez de los conocimientos computacionales, que facilitan el análisis de los datos, como por la limitación en lo referente a la cantidad de datos, puesto que resulta difícil que estos alcancen un nivel representativo, porque no son recopilados por un equipo sino por alguna persona en concreto. Además, otros investigadores carecen de los medios necesarios para tener contacto con los datos recopilados y así tener la posibilidad de reestudiarlos. Todo esto genera una pérdida innecesaria de recursos y a la vez se reduce la aplicabilidad de los resultados de la investigación en el resto del mundo.

Recientemente, muchas páginas web promocionan los conocimientos relacionados con la lingüística de corpus, así como con la creación de corpus, presentándolos como un elemento indispensable en los estudios lingüísticos y como algo que tienen que tener en cuenta todos los lingüistas. Nosotros opinamos que resulta de gran utilidad profundizar en la comprensión del mecanismo de creación de un corpus, puesto que se aprovecharían las ventajas procedentes de la interacción positiva entre las técnicas computacionales y los estudios de Humanidades.

Por otro lado, los estudios de interlengua desde los puntos de vista del Análisis Contrastivo y del Análisis de Errores se han centrado en la influencia de la lengua materna en la segunda lengua, que generalmente es el inglés, y se ha prestado menos atención a otros campos o idiomas. Como consecuencia de esto, queremos construir un corpus que se centra en poder ofrecer observaciones de la interacción entre diferentes interlenguas. Con este propósito, vamos a crear un corpus paralelo anotado procedente de aprendices, que estudian inglés como L2 y español como L3.

Para desarrollar estos contenidos, este trabajo va a tener la siguiente estructura: Se comenzará con el repaso de estudios precedentes en la sección 2. La sección 3 se dedica a la descripción de la motivación y el propósito de este estudio. En la sección 4, presentamos los procesos así como las herramientas que utilizamos para facilitar el procedimiento de creación. Por último, la sección 5 es la conclusión.

2. Estudios precedentes

2.1. Creación de corpus de aprendices

El propósito de crear nuevos corpus normalmente proviene de la necesidad de encontrar mecanismos y resultados que no pueden hallarse en los estudios precedentes. Esto nos conduce a repasar cuáles son los corpus que han aparecido en el mundo lingüístico con anterioridad. La mayoría de los corpus de aprendices conocidos se relacionan con la lengua inglesa, por ejemplo: Cambridge Learner Corpus (CLC), Hong Kong University of Science & Technology (HKUST), Corpus of Learner English, International Corpus of Learner English (ICLE), Longman Learners' Corpus (LLC)... etc.

Además de estos corpus, se conocen dos que recopilan datos de los aprendices que estudian español como segunda lengua: Corpus Escrito de Español como L2 (CEDEL2) y Spanish Language Learner Oral Corpus (SPLLOC). Sin embargo, la lengua materna de ambos corpus es inglés. El CEDEL2 está formado por 350.000 palabras y fue creado por la Universidad Autónoma de Madrid con el objetivo de estudiar el orden de las palabras (Proyecto de WOSLAC). Además cabe mencionar otro corpus importante llamado SPLLOC. Fue creado por la Universidad de Southampton, la Universidad de Newcastle y la Universidad de York y estudia principalmente clíticos como su meta.

Teniendo en cuenta estos datos, percibimos que el tema de los aprendices taiwaneses que aprenden español no ha sido algo que haya interesado a los investigadores de fuera de Taiwán, sin embargo, hoy en día existe la necesidad de abordar este tema tanto en lo referente al área de corpus como en lo que se refiere a la adquisición de la segunda lengua. Por eso hemos construido un corpus llamado CATE (Corpus de aprendices taiwaneses de español) que comenzó a desarrollarse en el año 2005. Este corpus ha recopilado 1.453 composiciones (aproximadamente 287.142 palabras) procedentes de 15 universidades de Taiwán en los últimos 3 años. La motivación de crear nuevo corpus procede de la necesidad de incluir datos de interlengua de L2, con la finalidad de investigar en el proceso de adquisición

sirviéndonos de un método más sistemático.

2.2. La interlengua

Persiguiendo el propósito, que hemos mencionado anteriormente, de crear un nuevo corpus para posibilitar el estudio de la interlengua de L2, no podemos desechar los estudios precedentes relacionados con este tema. De estos estudios anteriores vamos a combinar las teorías y las aplicaciones del Análisis de Errores y del Análisis Contrastivo, como base en el proceso de la creación de nuestro corpus de aprendices.

En primer lugar, Gass (2001: 79) sintetiza los pasos del Análisis de Errores: (1) recopilar datos, (2) identificar errores, (3) clasificar errores, (4) calcular errores, (5) analizar sus fuentes y (6) sugerir remedios. Por otro lado, Gass (2001: 72) indica que el Análisis Contrastivo supone lo siguiente: (1) la lengua materna es la fuente principal de los errores ocurridos en la producción de una segunda lengua, (2) los errores se pueden interpretar considerando las diferencias entre L1 y L2, (3) si las diferencias entre L1 y L2 son muchas, más errores se cometen, (4) lo que un estudiante tiene que aprender de una segunda lengua son las diferencias entre L1 y L2. Entonces, la construcción de nuestro corpus sigue los pasos (1)-(5) del Análisis de Errores en el siguiente orden: recopilamos los datos, identificamos los errores cometidos por los aprendices taiwaneses, clasificamos los errores según los diferentes tipos, analizamos sus fuentes en el proceso de anotación y calculamos los errores al consultar los datos utilizando la función de “corpus search”. Además, de acuerdo con el Análisis Contrastivo, no sólo comparamos L1 y L3 de nuestros aprendices, sino que recopilamos los datos de su interlengua de L2 para hacer las comparaciones posteriores.

En los estudios relacionados con la interlengua, Selinker (1972) propone que los factores que intervienen consisten en la transferencia, la sobregeneralización, la simplificación... etc. Con respecto a la transferencia, Taylor (1974) indica que la lengua materna juega un papel importante tanto en la transferencia positiva (o la facilitación en la terminología de Gass (2001)) como en la negativa (o la

interferencia). También pone de relieve que la lengua materna no es el único elemento que puede afectar al desarrollo de la interlengua, sin embargo no se puede ignorar su importancia. Centrándonos en los factores posibles, la transferencia de L1 y L2 se adopta en las etapas originarias de nuestra investigación como un sistema para anotar los errores estudiantiles, aunque probablemente adoptemos otras herramientas o mecanismos en un futuro.

Además, vale la pena prestar atención a lo que aportan Koike & Klee (2003) que mencionan que la transferencia es común y natural cuando dos lenguas se parecen mucho, y proporciona el ejemplo de un angloparlante que estudie japonés. Explica que como estas dos lenguas, inglés y japonés, tienen muchas diferencias entre sí, la transferencia no se produce tanto como en el caso de dos lenguas que comparten muchas similitudes. Koike & Klee (2003:32) añaden además que "saber cuáles son las estructuras que más se presentan en la transferencia a la L2, ayuda a prever muchos de los problemas con los que se enfrentan los estudiantes y a identificar la fuente más probable de los errores".

Si revisamos otras investigaciones en torno a este tema nos damos cuenta de que, en general, la mayoría de ellas se relacionan con el aprendizaje del inglés. Sin embargo, también encontramos algunos trabajos que se centran en la interlengua de español. Entre los cuales cabe destacar: Nieto & Martínez Vázquez (2006), Montrul (2000), Lozano (2006), Ramirez-Mayberry (1998)...etc. Si nos centramos en los aprendices taiwaneses, no se puede ignorar el estudio de Lin (1995).

Teniendo en cuenta lo mencionado anteriormente, tenemos que admitir la realidad de que los aprendices taiwaneses de español habían estudiado inglés antes que español. Por lo tanto, nuestro tema de investigación se centra en el caso de que haya más de una interlengua. Magiste (1979) señaló el problema de la competición en el sistema de multilenguas, o sea, llega a la conclusión de que la transferencia de LE1 a LE2 tiene un papel en la adquisición de multilenguas. Sin embargo, apuntó Gass (2001), la transferencia de la interlengua recibe muy poca atención en los estudios de Adquisición de Segunda Lengua (ASL), tal como lo apunta el estudio de De Angelis (2005).

A la luz de las aportaciones de estos estudios precedentes, nosotros planteamos la siguiente cuestión: ¿Es posible que L2 inglés, y no L1 chino, sea el que pueda ejercer un mayor grado de transferencia en L3 español para los aprendices taiwaneses que estudian inglés como su primera lengua extranjera y español como segunda lengua extranjera? En este sentido, sería interesante aportar datos que provengan de los aprendices taiwaneses cuya interlengua de L2 es inglés y aprenden español como L3. Por ello, la creación de un corpus paralelo de aprendices puede ser un recurso para verificar la hipótesis. En nuestro trabajo anterior (Lu & Lu) (2008)) anotamos los errores estudiantiles de 10 aprendices principiantes de español utilizando los sistemas de L12, L23 y L13 con rasgos de S (Sí, hay influencia) o N (No, no hay influencia)[3]. El trabajo llegó a las siguientes conclusiones: (1) en el caso de los aprendices principiantes que llevaban 160 horas estudiando el español, la influencia de L2 afecta más que L1 en su producción escrita. (2) Los errores causados por la influencia de L1 se asocian más con la flexión verbal de tiempo y con las preposiciones. (3) Los tipos de errores afectados por L1 tienden a relacionarse con la ortografía, la concordancia con los sustantivos y las preposiciones.

3. Motivación y propósito de la creación de CPATEI

La motivación de crear un corpus proviene de una necesidad actual y de una meta futura: la falta de recursos similares en el presente, la necesidad de compartir recursos y la facilitación de nuevos datos de consulta en el futuro. Los corpus creados hasta este momento no satisfacen nuestra necesidad de poder comparar L2 (inglés) y L3 (español) en el mismo corpus, por ejemplo, al CATE le falta la posibilidad de realizar un contraste entre diferentes interlenguas, mientras que los

[3] En la primera etapa, anotamos los errores con L12, L13, L23 y LIE. Las anotaciones de L12, L13 y L23 se refieren a las interrelaciones entre las tres lenguas que han estudiado los alumnos, por ejemplo, la relación entre el chino y el inglés, el chino y el español y el inglés y el español. La anotación de LIE se refiere al uso correcto o incorrecto en las composiciones escritas en español y las traducciones en inglés. A través de un proceso de examen de los resultados anotados relacionados con las tres lenguas, el chino, el inglés y el español, podemos deducir cuál es el factor principal de los errores, L1 o L2.

demás corpus no tienen nada que ver con los aprendices taiwaneses cuya lengua materna es el chino-mandarín[4].

Nuestro propósito final consiste en desentrañar los procesos de aprendizaje en el supuesto de interferencia de más de una interlengua mediante la creación de corpus, porque se trata de un campo de estudio en ASL que todavía requiere mucha dedicación. Esperamos que la construcción de este corpus paralelo de aprendices taiwaneses nos facilite análisis futuros que nos permitan profundizar en los estudios de ASL referentes a la interacción de distintas interlenguas.

4. Construcción del corpus: Corpus paralelo de aprendices taiwaneses de español e inglés (CPATEI)[5]

Después de establecer el propósito y el criterio que se va a adoptar en lo concerniente al diseño del corpus, en el procedimiento de creación, recopilamos los datos, los archivamos y los anotamos usando las herramientas apropiadas.

4.1. Recopilar, escribir a máquina, archivar los datos y corregir las composiciones

Cabe mencionar, en primer lugar, que este es el primer año en el que construimos CPATEI, por eso los datos que hemos recogido no conforman una cantidad ideal. En este corpus participan 29 alumnos del Departamento de Lenguas Extranjeras de la Universidad Nacional de Cheng Kung. Estos estudiantes poseen características diferentes: hay 12 alumnos que son del segundo

[4] Se refiere a la comparación entre los corpus CATE y CPATEI con motivo de conocer la diferencia que hay entre ellos. En primer lugar, CATE se ha construido desde el año 2005 y ahora contiene 287.142 palabras. Mientras tanto, CPATEI está recientemente construido y ahora está formado por 6.735 palabras en español y 7.136 palabras en inglés. En segundo lugar, en CATE se recopilan sólo las composiciones escritas en español, sin embargo, en CPATEI, hay tanto composiones escritas en español como las traducciones en inglés. En tercer lugar, se anotan los errores corregidos por los hablantes nativos en ambos corpus. Pero en CPATEI, además se anotan los factores de la transferencia de L1 y L2. Por último, CATE se consulta a través de la interfaz del internet, mientras tanto, CPATEI se consulta mediante diferentes herramienta auxiliares.

[5] La amplificación de CPATEI se basa en la investigación que hemos realizado con motivo del Noveno Simposio sobre la Didáctica, la Cultura y la Traducción de Español, que tuvo lugar el día 3 de mayo en el Departamento de Español de la Universidad Tamkang. En este corpus, hemos añadido 19 composiciones escritas en español y sus traducciones en inglés a las 10 composiciones incluidas anteriormente. Esperamos que los resultados del estudio puedan servir como sugerencias para los investigadores que tengan un interés similar.

curso y han estudiado español 160 horas. Mientras tanto, hay 17 alumnos que son del tercer curso y ellos han estudiado español 320 horas. Entre ellos, hay 28 alumnas y un alumno. Esto es un reflejo de la situación actual en Taiwán, donde hay más chicas que chicos que aprenden lenguas extranjeras. Además, suelen empezar a aprender español como lengua extranjera en la universidad.

Los datos recopilados provienen de dos partes: de las composiciones escritas en español y de las traducciones de estas composiciones al inglés. El tema de la composición del segundo curso es la presentación de sus amigos, mientras que el de la composición del tercero es contar los problemas que tenían cuando eran adolescentes. Ambos trabajos se han terminado fuera de la clase.[6]

Después de recopilar los datos, los pasamos al ordenador, luego los revisamos y archivamos en formato de texto con codificación de Unicode UTF-8[7], para guardar algunas letras y puntuaciones especiales del español. De momento el corpus está compuesto de 6.735 palabras en español y 7.136 en inglés, respectivamente.[8]

Al guardar y archivar los datos, repartimos las composiciones en español y las traducciones en inglés a dos hablantes nativos[9] para que nos ayuden a corregir los errores cometidos por los aprendices.

4.2. Anotación con Corpus Tool[10]

Utilizamos Corpus Tool, sistema diseñado por la Universidad Autónoma de Madrid (UAM) puesto que lo consideramos una herramienta útil para realizar

[6] Teniendo en cuenta la experiencia anterior en lo referente a la construcción del corpus, en el futuro vamos a establecer unas reglas fundamentales: que los alumnos deberían escribir las composiciones en español y las traducciones en inglés dentro de la clase y que el tema de la composición debería ser el mismo para todos los participantes con el fin de controlar las variables que podrían afectar a los resultados finales.

[7] Agradecemos a Ciro Chang que ha diseñado un programa informático para transformar los datos escritos de forma document a la de text.

[8] La información estadística de los textos estudiantiles se muestra en el siguiente cuadro.

	Español	2do curso	3er curso	Inglés	2 do curso	3 er curso
Cantidad de palabras/texto	232	214	245	246	228	256
Cantidad de oraciones/texto	19	23	16	19	23	16
Promedio de letras/palabra	4,6	4,6	4,6	4,3	4,3	4,3
Promedio de palabras/oración	12,9	9,6	15,2	13,8	10,4	16,2

[9] Querríamos dar nuestro agradecimiento a María H. Contreras y Derek Murphy, los dos hablantes nativos que nos ayudan a corregir las composiciones en español y las traducciones en inglés.

[10] Agradecemos a UAM por permitirnos usar de forma gratuita el Corpus Tool.

anotaciones. Esta herramienta es como una programación de editor y entre sus características cabe destacar que no sólo ahorra tiempo en teclear los signos de anotación y evita la probabilidad de cometer errores propios de los seres humanos, sino que también facilita la combinación de rasgos de diferentes sistemas bajo una interfaz conveniente. Además, los resultados anotados y etiquetados pueden guardarse en el formato de XML, que es compatible con las demás herramientas de corpus que utilizamos en las siguientes fases de creación. En definitiva, adoptamos esta herramienta para anotar las correcciones correspondientes a los errores, los tipos de errores y la transferencia de L1 o L2 en L3. Los datos están incorporados al proyecto de Corpus Tool en dos subcorpus: uno compuesto de las composiciones en español y otro con las traducciones en inglés.

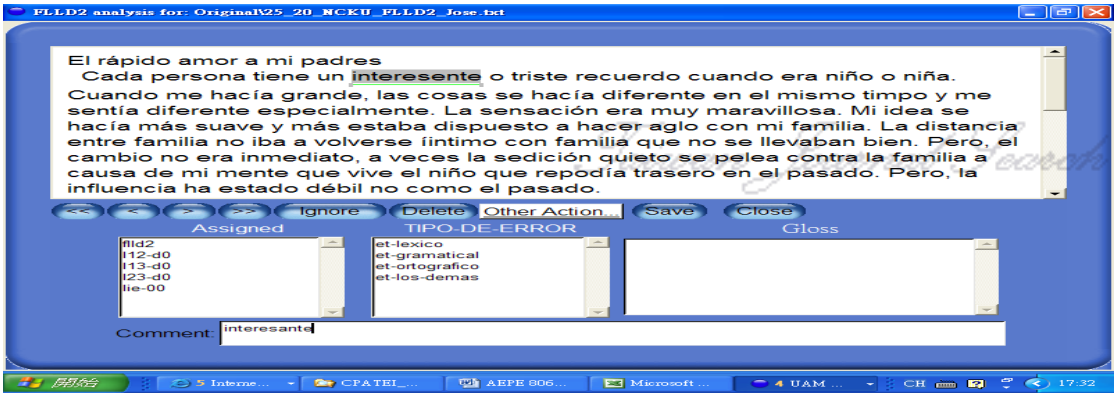
En lo referente al trabajo de anotación, hemos tomado determinadas medidas para controlar la calidad del corpus, en primer lugar, un anotador toma nota de los errores con correcciones correspondientes o asigna un rasgo elegido de diferentes sistemas tales como tipo de error o transferencia. Después, un segundo anotador revisa cuidadosamente todas las decisiones tomadas por el primero. Por último, si existe desacuerdo entre los anotadores anteriores, un tercero toma la decisión final considerando las anotaciones hechas por los dos anotadores anteriores.

4.2.1. Anotación de las correcciones correspondientes a los errores cometidos[11]

En primer lugar, hay que señalar que nos hemos servido de la función "comment" de Corpus Tool para anotar las correcciones correspondientes a los errores cometidos por los aprendices. El procedimiento que seguimos incluye varios pasos, primero marcamos mediante un proceso de subrayado los errores en las composiciones originales, y luego escribimos las correcciones en el espacio titulado "comment" (véase Figura 1).

Figura 1

Anotación de los errores con las correcciones correspondientes



Las correcciones se clasifican en tres tipos, como se muestra en el Cuadro 1. Hemos considerado que es mejor que no se deje ningún espacio en blanco y que se anote "000" cuando el tipo de corrección corresponde a una omisión de los elementos escritos por los aprendices. Aunque esto requiere más un mayor grado de dedicación en lo que se refiere a la programación que se utiliza para convertir las composiciones originales en textos corregidos, merece la pena hacerlo porque puede disminuir la probabilidad de cometer algún error por parte de los anotadores en el proceso de creación.

[11] Anteriormente los errores están anotados a mano utilizando el formato XML. Sin embargo, nos servimos de la herramienta de Corpus Tool para reproducir directamente los resultados anotados con el formato XML.

Cuadro 1

Clasificaciones de corrección

Clasificaciones de corrección	Subrayar los errores cometidos por los alumnos		Anotar correcciones correspondientes a errores estudiantiles en “comment”
Sustituir	xxx yyy zzz	→	aaa
Omitir	xxx yyy zzz	→	000
Insertar	xxx_zzz (insertar una raya entre dos palabras)	→	aaa

Además, cuando anotamos la corrección del tipo traslación (cambiar un elemento de una posición original a otra distinta), adoptamos el método de anotación de multiniveles: anotamos todas las correcciones en el segmento largo causado por la traslación, incluyendo todos los elementos menores corregidos. Esto lo hacemos así por dos razones: (1) para que la programación de texto corregido pueda distinguir los segmentos diferentes fácilmente y (2) para que el cálculo no sea erróneo debido a los elementos repetidos[12].

Por último, queremos mencionar que la utilización de “comment” no sólo nos es de utilidad en el sentido de elegir uno entre los demás tipos de anotaciones, sino que también facilita el proceso de convertir los errores originales en las correcciones anotadas de los textos exportados con el formato de XML.

4.2.2. Anotación de tipos de errores y transferencia[13]

Hemos diseñado un sistema titulado “tipo de error” con 4 rasgos fundamentales: gramatical, léxico, ortográfico y otros, que nos sirve para clasificar los tipos de

[12] Si, para anotar una traslación, en vez de seleccionar el tipo “sustituir” utilizamos los tipos “omitir” el elemento original e “insertar” el elemento omitido, un mismo error se va a calcular dos veces y además va a asignarse un tipo de error incorrecto.

[13] En una etapa posterior a la primera versión y anterior a esta versión, anotamos los errores teniendo en cuenta los factores D (si las dos lenguas son Diferentes) o S (si las dos lenguas son Similares). Después de evaluar las desventajas de los dos tipos de anotación (SN en la versión anterior y DS en la versión posterior), decidimos adoptar el término “transferencia” como un sistema nuevo para anotar los errores según los factores que afectan a la interlengua.

errores cometidos por los aprendices (véase Cuadro 2).

Cuadro 2

Esquemas y Rasgos

Esquema	Rasgo
Tipo de error	Léxico
	Gramatical
	Ortográfico
	Otros
Transferencia	L1
	L2-correcto
	L2-incorrecto
	No

De esta manera, en el mismo nivel que el sistema “tipo de error”, situamos otro sistema, el de “transferencia”, la cual que posee cuatro rasgos: “transferencia de L1, transferencia de L2, transferencia de interlengua de L2 y no-transferencia”. Para facilitar la consulta de los resultados considerando la complejidad de diferentes rasgos, se combinan los dos sistemas en el mismo nivel, en vez de separarlos en dos distintos motivos relacionados con la programación.

4.3. Consulta de tipo de error y tranferencia

Para consultar los resultados de los errores anotados, utilizamos la función de “corpus search” de Corpus Tool. Esto lo hacemos con distintos propósitos de investigación, puesto que nos permite obtener los resultados teniendo en cuenta diferentes tipos de errores o aunando dos clases de sistemas (tipo de error y transferencia), así como combinando cierto tipo de error con cierto tipo de transferencia.

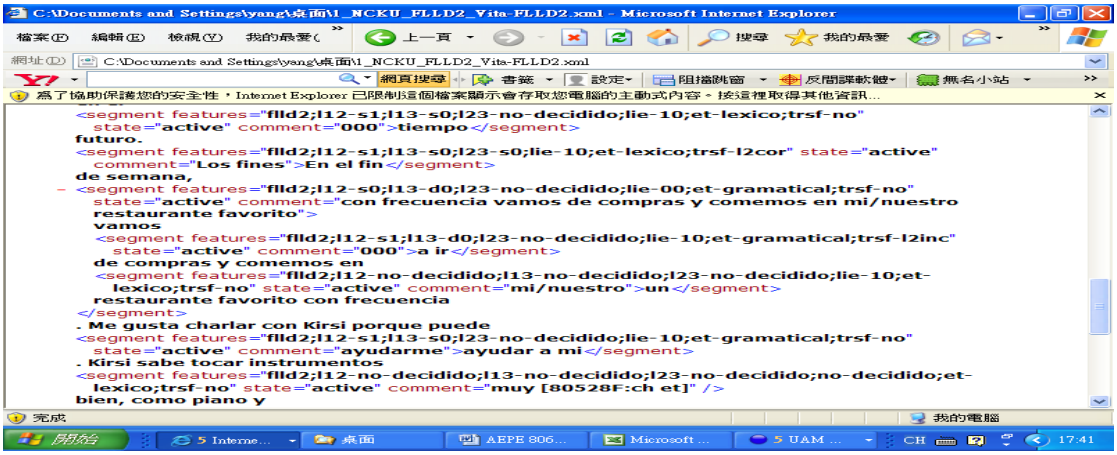
4.4. Textos corregidos

Además de las funciones incluidas en el Corpus Tool tal como “corpus search”, también tratamos de combinar los resultados exportados con otras técnicas computacionales, con el fin de hacer nuestro corpus más relevante en el marco de la investigación. Por ejemplo, los textos con correcciones anotadas en formato XML, se pueden convertir en los textos corregidos para facilitar los siguientes pasos que implica el desarrollo de un corpus de multiusos.

Las composiciones anotadas se muestran en Figura 2. Resulta útil explicar un poco las anotaciones. En este sentido, tenemos que señalar que yyy es todo lo que contiene una anotación. significa el comienzo de una anotación y se entiende como el fin de una anotación. Por ejemplo, en el caso de yyy, yyy son los errores cometidos por los aprendices. Mientras tanto, en , significa que “xxx” es la corrección correspondiente al error corregido por los hablantes nativos. Si aparece la anotación , “000” quiere decir que estas palabras deben omitirse. Si aparece únicamente , indica que necesita añadirse algo para completar el significado de la frase escrita por los aprendices y hacerla gramatical. En lo concerniente a la omisión, ya se ha mencionado su importancia y forma de análisis en la sección 4.2.1 Sin embargo, para conseguir omitir los elementos insertados 000, necesitamos añadir elemento caligráfico en el proceso de programación.

Figura 2

Anotado texto exporatdo con el formato de XML


C:\Documents and Settings\yang\桌面\1_NCKU_FLLD2_Vita-FLLD2.xml - Microsoft Internet Explorer
tiempo
futuro.
En el fin
de semana,

vamos
a ir
de compras y comemos en
un
restaurante favorito con frecuencia

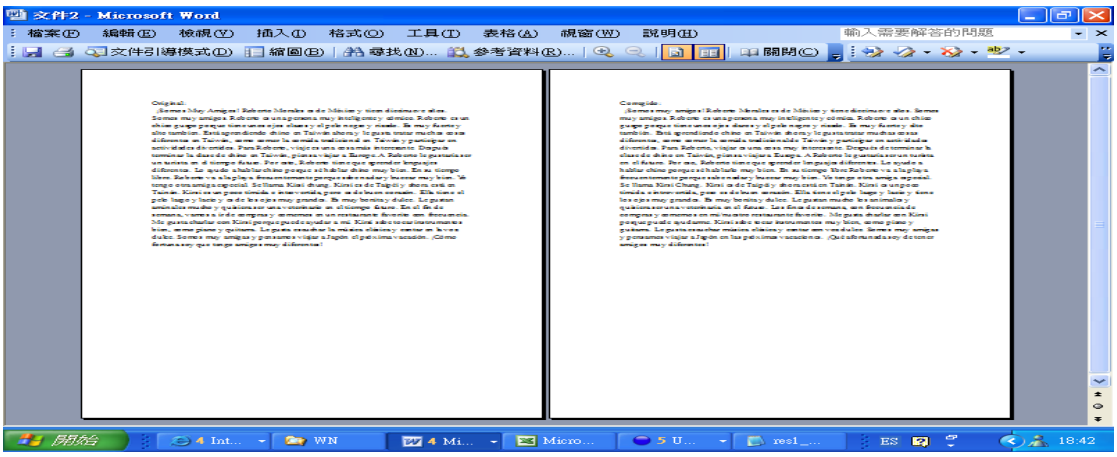
. Me gusta charlar con Kirsi porque puede
ayudar a mi
. Kirsi sabe tocar instrumentos

bien, como piano y

Además, hay que aclarar algunos puntos referentes a las correciones relacionadas con multiniveles. Las anotaciones que asignamos son distintas si se trata del segmento largo o de segmentos cortos, puesto que el más largo contiene todas las correcciones que poseen los segmentos cortos. Esto implica la necesidad de realizar un paso más para facilitar la programación computacional. En el texto con formato XML, el segmento más largo se distingue mediante el signo “-” antes que en el signo “<”, de manera que nuestro experto computacional sólo pone atención en este segmento ignorando los cortos para no repetir las correcciones y causar la agramaticalidad oracional. En el proceso de creación se produjo un continuo intercambio de ideas entre nosotros y los expertos y la realización de numerosas pruebas contemplando diversas posibilidades; de forma que al final elegimos esta opción como la más adecuada puesto que no interfiere en el cálculo correcto de los errores (véase Figura 3).

Figura 3

Texto original vs. texto corregido



4.5. Etiquetación de POS

Después de obtener las versiones corregidas por los hablantes nativos, podemos anotar el contenido de las composiciones teniendo en cuenta las partes de la oración (POS) con la finalidad de aumentar la utilidad del corpus, puesto que añadiendo información relacionada con POS aporta un mayor valor consultivo a este corpus. Para etiquetar los textos revisados con POS nos servimos de la herramienta Tree Tagger en vez de etiquetar los textos originales de los estudiantes. Por último, queremos señalar que el proceso de etiquetado POS no sólo se ha aplicado a los textos en español sino también a los que están en inglés.

4.6. ParaConc y datos paralelos

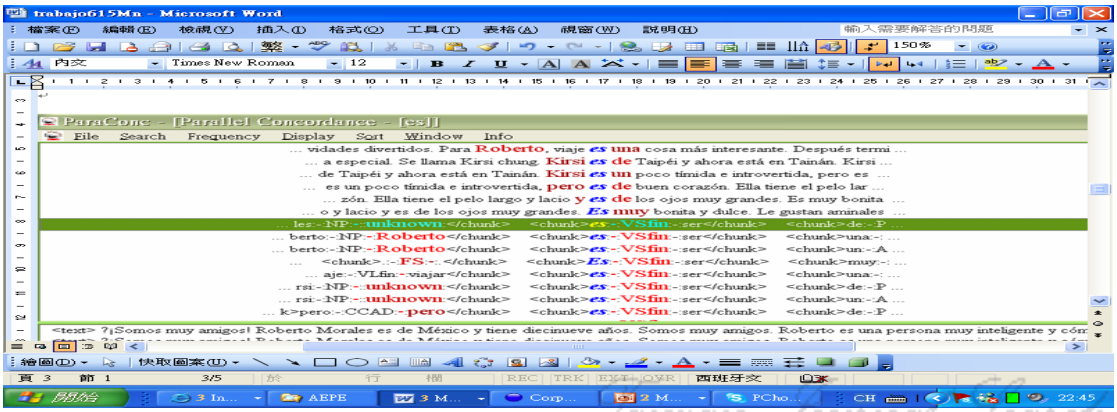
Al terminar la etiquetación de POS para ambos tipos de textos (en español y en inglés), nos servimos de la herramienta ParaConc para consultar los resultados contrastivos de los textos. ParaConc es una herramienta que está diseñada para realizar comparaciones y contrastes entre textos paralelos.

Antes de importar los textos a la herramienta, tenemos que alinear las oraciones de los textos paralelos para hacerlas comparables. La función paralela se aplica en dos sentidos en el presente estudio: el que se refiere a los distintos tipos de textos dentro de la misma lengua y el que se relaciona con los distintos textos en diferentes

idiomas. Al establecer contrastes entre los textos originalemente escritos por los aprendices y los textos corregidos por los hablantes nativos, nos ayuda a distinguir las diferencias entre la interlengua y la lengua meta (véase Figura 4).

Figura 4

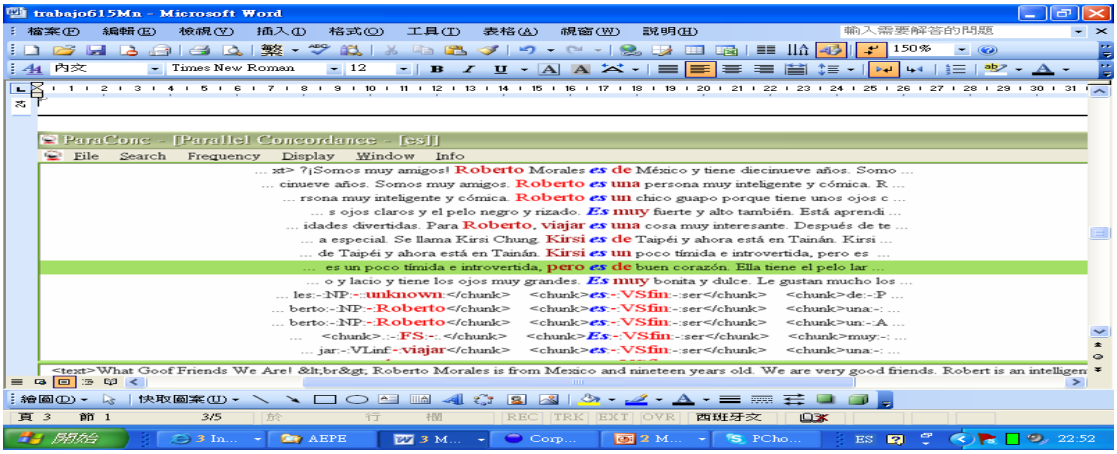
ParaConc: POS-etiquetado texto original vs. POS-etiquetado texto corregido



Además este proceso de contraste nos ayuda a distinguir detalladamente las diferencias entre dos lenguas diferentes, en este caso el inglés y el español. (véase Figura 5).

Figura 5

ParaConc: POS-etiquetado texto corregido en español vs. POS-etiquetado texto corregido en inglés



5. Conclusión

Con el fin de facilitar el estudio de los aprendices taiwaneses cuya lengua materna es el chino-mandarín y cuya primera lengua extranjera es el inglés, en este trabajo, nos hemos dedicado a la creación de CPATEI con el propósito de compartir la experiencia de construir corpus de aprendices. Hemos tratado de aplicar diferentes herramientas que hemos probado durante el proceso de construcción, para aportar visión global de los recursos disponibles. En este sentido queremos manifestar que estas herramientas son recomendables no sólo para usuarios sino también para profesionales e investigadores, que además podrían enriquecer su trabajo con las técnicas computacionales oportunas.

Cabe destacar, por último que el proceso de creación de corpus es un trabajo duro que requiere una dedicación constante y un esfuerzo continuo por parte de un equipo de investigadores. Pensamos que todo este esfuerzo merece la pena porque muchos investigadores y muchos más aprendices podrán beneficiarse en el futuro si el corpus sigue construyéndose conforme a un proceso bien preparado y diseñado.

Referencias Bibliográficas

De Angelis, G.. “Interlanguage transfer of function words.” *Language Learning* 99.3 (2005): 379-414.

Gass, S. & L. Selinker. *Second Language Acquisition: An Introductory Course*. London: Lawrence Erlbaum, 2001.

Koike, D. & C. Klee. *Lingüística Aplicada*. New York: John Wiley & Sons, 2003.

Lin, Y-H. Un “Análisis Empírico de la Estabilización/fosilización: La Incorporación y la Auto-corrección en Unos Sujetos Chinos.” Tesis doctoral, Universitat de Barcelona, 1995.

Lu, Hui-Chuan & Lu Lo Hsueh. “Estudio de interlengua a partir de un corpus anotado y paralelo entre el inglés y el español.” Ponencia presentada en el Noveno Simposio sobre la Didáctica, la Cultura y la Traducción de Español en la Universidad de Tamkang, 2008.

Lozano, C. “La adquisición del español como L2: la interfaz sintaxis-discurso.” Actas del XXIII Congreso Internacional de la Asociación Española de Lingüística Aplicada. Eds. Amengual Pizarro, M., M. Juan Garau & J. Salazar Noguera. Barcelona: Universitat de les Illes Balears (2006): 95-100.

Magiste, E. “The competing language systems of the multilingual: A developmental study of decoding and encoding processes.” *Journal of Verbal Learning and Verbal Behaviour* 18 (1979): 79-89.

Montrul, S. “Transitivity alternations in L2 acquisition.” *SSLA* 22 (2000): 229-273.

Nieto, H & J. Martínez Vázquez. “Análisis de caso: Evaluación de los resultados del taller de escritura del nivel avanzado de español para extranjeros.” *Revista Electrónica de Didáctica/Español Lengua Extranjera* 7 (2006):1-15.

Ramirez-Mayberry, M. “Acquisition of Spanish definite articles by English-speaking learners of Spanish.” *Texas Papers in Foreign Language Education* 3.3 (1998): 52-67.

Selinker, L. “Interlanguage.” *International Review of Applied Linguistics* 10 (1972): 209-230.

Taylor, B. P. “Toward a theory of language acquisition.” *Language Learning* 24.1 (1974): 23-35.

Herramientas:

Corpus Tool:

Universidad Autónoma de Madrid. 2008. Corpus Tool.

<http://www.wagsoft.com/CorpusTool/>

ParaConc:

Barlow, Michael. 2004. ParaConc.

Cambridge Learner Corpus (CLC)

Corpora de aprendices:

CEDELS2:

Corpus Escrito de Español como L2 (CEDEL2)

http://www.uam.es/proyectosinv/woslac/cedel2.htm

CLC:

Cambridge Learner Corpus (CLC)

http://www.cambridge.org/elt/corpus/learner_corpus.htm

http://www.ust.hk/eng/index.htm

ICLE:

Corpus of Learner English, International Corpus of Learner English (ICLE)

http://www.fltr.ucl.ac.be/fltr/germ/etan/CECL/Cecl-Projects/Icle/icle.htm

LLC:

Longman Learners' Corpus (LLC)

http://www.pearsonlongman.com/dictionaries/corpus/learners.html

SPLLOC:

Spanish Language Learner Oral Corpus (SPLLOC)

http://www.splloc.soton.ac.uk/