EL ESTUDIO DE LAS ETAPAS DEL APRENDIZAJE LÉXICO BASADO EN EL CATE-CIC

Hui-Chuan Lu
National Cheng Kung University

Resumen: Tomando como base el Corpus de Aprendices Taiwaneses de Español, creado por Lu, y utilizando los instrumentos analíticos específicos de corpus, hemos pretendido estudiar el desarrollo léxico de los estudiantes que aprenden el español como segunda lengua extranjera.
Nos hemos centrado en el cotejo de las palabras combinatorias empleadas en distintos niveles de aprendizaje y propuesto un continum en lo que se refiere al grado de facilidad del proceso de aprendizaje, que sería el siguiente: V-Prep-V/N-Prep-N, V-Conj, V-Adv/V-N, N-Adj, V-Adj, Adj-N, N-V, Det-N. Además, no sólo nos percatamos de que en las etapas de aprendizaje, existe un orden de facilidad a dificultad que va desde el tipo otorgráfico hasta el gramatical y el léxico; sino que también especificamos que la dificultad del sentido semántico aparece en una etapa más tardía que el vocabulario con respecto al tipo léxico de las palabras combinatorias.

Palabras clave: corpus, léxica, interlengua.

Abstract: *Taking the base of the Corpus of Taiwanese Learners of Spanish, created by Lu, and applying as special analytical instruments, we've tried to study the lexical development of the students learning Spanish as second foreign language.*
We've focused on the comparison of the combination of words used in different levels of learning in which we proposed a continuum for the degree of learnablity form the easiest to the most difficult as follows V-Prep-V/N-Prep-N, V-Conj, V-Adv/V-N, N-Adj, V-Adj, Adj-N, N-V, Det-N. Moreover, we not only derived, in the developmental stages of learning, an order from early stage to late stage exists that comes from the orthographic type to the grammatical and then to lexical types; but we also identified that the semantic meaning appears in a stage later than the vocabulary with regard to the lexical type of the combinative words.

Key Words: *corpus, lexical, interlanguage.*

1. INTRODUCCIÓN

Los corpus se han convertido en una fuente importante para casi todos los estudios relacionados con la lingüística, aunque su influencia y los resultados posteriores pueden ser diferentes dependiendo de los idiomas a los que se refiera. El presente persigue dos objetivos fundamentales: por una parte, la construcción del primer corpus de los estudiantes que aprenden otra lengua extranjera diferente al inglés; y por otra parte, el estudio de las etapas del desarrollo de la interlengua para el campo léxico de la lingüística. El propósito de combinar la construcción de un corpus de los estudiantes y su aplicación en el análisis léxico es demostrar cómo puede ser eficaz y de gran alcance la metodología basada en el corpus para los estudios de la lingüística aplicada.

El desarrollo de los corpus relacionados con los idiomas extranjeros, con excepción de inglés, ha sido lento de forma que el número de corpus españoles existentes es muy limitado;

dos ejemplos serían el CEDEL2 (Corpus Escrito del Español L2) y el corpus de los estudiantes japoneses del español. Sin embargo, un 21.4% de las investigaciones que se relacionaron con la lingüística aplicada se realizaron en Taiwán a partir de las redacciones de los estudiantes recogiendo y usando datos individualmente, pero sin considerar la posibilidad de uso público (de tales recursos valiosos) mediante una puesta en común. Por lo tanto, desde nuestro punto de vista, existe la necesidad imperiosa de la creación del corpus de los estudiantes taiwaneses del español (CATE: Corpus de Aprendices Taiwaneses de Español).

Además, basándonos en el análisis de los datos extraídos del corpus elaborado por los estudiantes y teniendo en cuenta la frecuencia de aparición y examinando las relaciones estadísticas, demostraremos cómo un corpus puede ser de gran utilidad para arrojar luz sobre las etapas del desarrollo de diversos niveles de estudiantes en diversos subcampos de la lingüística, tales como el léxico (la colocación y la coligacón). En definitiva, la aportación principal de este trabajo será la creación del primer corpus de los estudiantes taiwaneses del español. Además, proponemos modelos de análisis de los datos extraídos del corpus construido con el fin de proporcionar un muestrario objetivo y representativo del desarrollo de la adquisición de la lengua por parte de los estudiantes.

2. ESTUDIOS PRECEDENTES

2.1. Construcción de corpus de estudiantes

En lo concerniente a los estudiantes del inglés, cabe señalar la existencia de los siguientes corpus: el ICLE (Corpus internacional del inglés de los estudiantes), el de Longman, el de Cambridge[1] y Montclair[2] (la base de datos electrónica de los alumnos de la lengua de Montclair). En Taiwán, existen los siguientes corpus de los estudiantes de la lengua inglesa: el constituido por Coleman Bernath en la Universidad Soochow, el de NTOU (el Corpus de los estudiantes taiwaneses EFL, que contiene 53.000 palabras y fue constituido por Hau-ran Chen en NTOU), y el TLCE constituido por la profesora Shih de la Universidad Nacional Su-Yet-Shen (se trata del corpus más representativo de Taiwán de los de estudiantes de inglés y está basado en un sistema de etiquetas). Además, ENGLISHTLC[3], se trata de un corpus inglés electrónico de los estudiantes taiwaneses con un sistema que puede detectar los errores. En China, encontramos el corpus de los estudiantes chinos del inglés[4] (1.000.000 de palabras), creado por Gui Shichun del departamento del idioma extranjero en la Universidad Guang Dung. Además, existe el corpus MET (400.000 palabras) que fue constituido por He Anping de la Universidad Meridional Nacional de China, así como el corpus del inglés de los estudiantes[5] (25.000.000 de palabras) fundado por J. Milton, que contiene etiquetas de las categorías gramaticales y de los errores.

Si atendemos a otros idiomas extranjeros diferentes al inglés, el desarrollo de los corpus de los estudiantes japoneses[6] ha llamado mucho más la atención en Asia, por ejemplo, el corpus de los estudiantes japoneses creado por la Universidad Nacional Cheng Kung en Taiwán. Sin

[1] http://leo.meikai.ac.jp/~tono/lcorpuslist.html
[2] http://chss.montclair.edu/linguistics/MELD/
[3] http://mail.tku.edu.tw/dwible/index.htm
[4] http://nora.hd.uib.no/corpora/1997-1/0262.html
[5] http://leo.meikai.ac.jp/~tono/lcorpuslist.html
[6] http://140.116.245.232/japan_corpus

embargo, encontramos un número muy reducido de corpus establecido por estudiantes de la lengua española, por ejemplo: (1) el Corpus Escrito del Español L2 (CEDEL2)[7] que recopila trabajos escritos de los principiantes que aprenden el español como segunda lengua, y (2) un sub-corpus del Corpus Internacional de los estudiantes de inglés-ICLE creado por JoAnne Neff de la Universidad Complutense en Madrid, y (3) Corpus japonés de los estudiantes del español[8] .

2.2. Estudios a partir de los corpus de estudiantes

Entre los estudios generales basados en los corpus de los alumnos podemos mencionar las aportaciones de los siguientes autores: Granger (1998), Neff (2004) y Mönnink et al. (2000), entre otros. Por ejemplo, Mönnink *et al*. analiza los errores de la sintaxis de los estudiantes usando el sistema de las etiquetas de ICLE. Con respecto al análisis de los datos extraídos de los corpus de los estudiantes chinos del inglés cabe mencionar: Huang (2000), Chuang (1996), Wang (1999), Flowerdew (2006), etc. Por ejemplo, Flowerdew trata de las clasificaciones de los errores basadas en los corpus de los estudiantes cantoneses del inglés. Estos estudios precedentes proporcionan una conclusión generalizada y representativa mediante el análisis pormenorizado de una gran cantidad y variedad de datos.

2.3. Estudios relacionados con la interlengua

En lo referente a las investigaciones relacionadas con la interlengua, Kitsnki (2006) propone que el análisis basado en los corpus de los estudiantes y en los estudios de la interlengua puede ayudar a detectar los tipos de errores así como las características de la interlengua desarrollada por los estudiantes. Cobb (2003) prueba que la recopilación de datos basados en los corpus facilita el resultado de la investigación en el aprendizaje de segundas lenguas. Liang (2006) aplica las etiquetas de las categorías gramaticales de CLAW en el análisis de las redacciones de los estudiantes chinos del inglés para reforzar el estudio de la interlengua. Además, Lisi (2001) investiga las etapas desarrolladas por los estudiantes españoles del inglés a partir de un corpus que contiene una gran cantidad de datos. También, Gonzalez-Espresati (2002) exploró la interlengua comparando y analizando errores cometidos por los estudiantes españoles del portugués y los estudiantes portugueses del español. Por otra parte, Alexopoulou (2005) analiza la interlengua para construir errores sintácticos de la expresión escrita de los estudiantes españoles del griego. Además, Neff et al. (2003) compara el inglés de los hablantes nativos y el de los estudiantes e indica que los estudiantes usan demasiado o demasiado poco los verbos auxiliares. Aijmer (2004) compara las diferencias de los usos de la interlengua entre los estudiantes suizos y los hablantes nativos del inglés. Se asume que la razón principal de los errores que cometen los estudiantes tiene que estar relacionada con la lengua materna de los estudiantes. Por ejemplo, el error de la conjunción inglesa se relaciona con la lengua nativa de los estudiantes (Tseng y Liou, 2006). Según el análisis de los datos en los corpus de los trabajos escritos de los estudiantes árabes del inglés y según el Corpus Louvain de ensayos de los nativos ingleses (Al-Btoosh, 2005) podemos decir que los trabajos escritos están más cerca del uso de su lengua natal.

[7] http://www.uib.no/mailman/public/corpora-archive/2006-June/002850.html
[8] Y. Kamakura (2005)

3. CATE

3.1. CATE 2005-2006

3.1.1. La estructura

En primer lugar, queremos agradecer el apoyo financiero de NSC (NSC93-2411-H-006-022, NSC 94-2411-H-006-016 y NSC95-2411-H-006-012) que nos ha facilitado la creación del corpus. Además, damos las gracias especialmente a los estudiantes que nos han proporcionado sus redacciones, así como a los profesores de la Universidad católica de Fu Jen, la Universidad de Tamkang, la Universidad de Providence, el Colegio universitario de Wenzao Ursuline de los idiomas, la Universidad Nacional de Taiwán, la Universidad Nacional de Chengchi, la Universidad de NanHua, la Universidad China de Ultramar de la Tecnología, la Universidad Nacional de Formosa, la Universidad Nacional de Chiayi, y la Universidad Nacional de Cheng Kung por ayudarnos a recoger los trabajos[9].

En nuestros planes futuros, con la intención de ampliar el CATE (Corpus de Aprendices Taiwaneses de Español), seguiremos recopilando datos de los estudiantes taiwaneses del español mediante un procedimiento que consistente en recoger, corregir, introducir y programar los datos de las redacciones. Ampliaremos sistemáticamente la cantidad del corpus año tras año por medio de la recogida de 500 a 1000 redacciones cada año para conseguir el objetivo final que pretendemos: que el CATE sea el corpus más grande construido en Taiwán.

Con vistas a compartir estos recursos, todos los autores de las redacciones han firmado las cartas de consentimiento y han estado de acuerdo en que sus trabajos se pudieran estudiar por el público en general en un futuro.

En cuanto al criterio de creación, el CATE sigue algunos principios de la construcción de un corpus, pretendiendo ser representativo, tener diversidad y tener balance. Actualmente, hemos acabado el trabajo estructural de los primeros dos años (2005-2006).

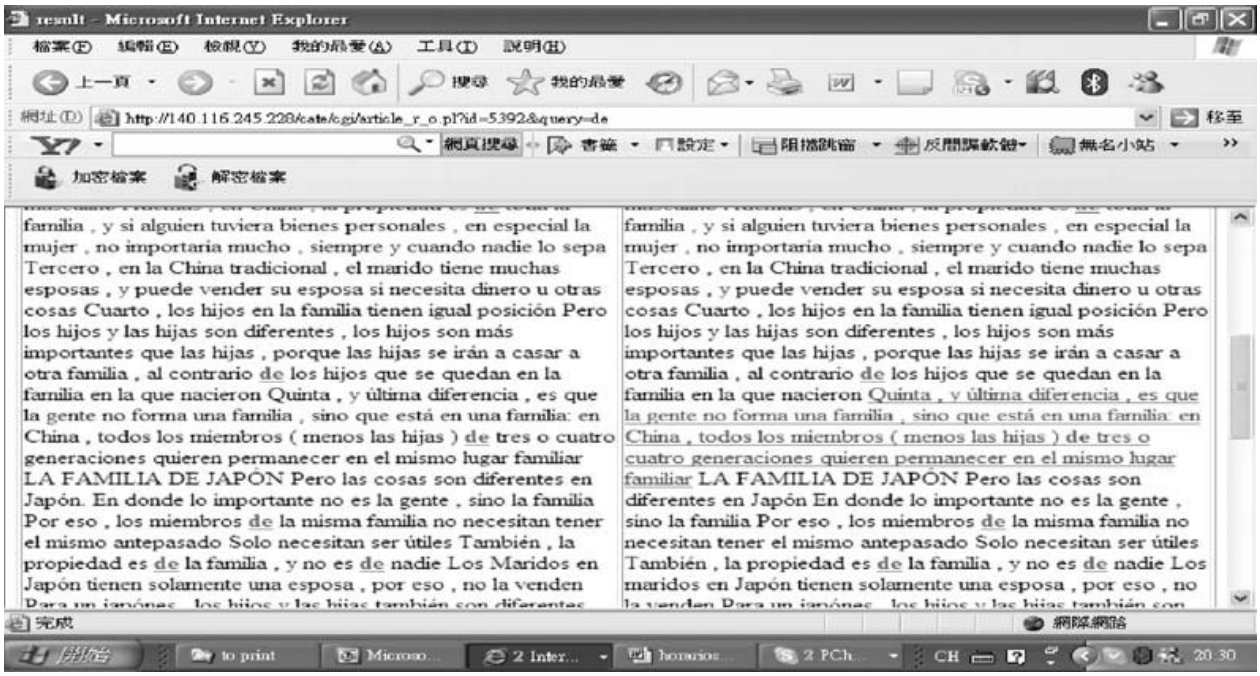
Las fuentes de las que proceden los datos del corpus consisten en dos grupos de estudiantes: (1) en nuestro primer grupo, los estudiantes del departamento del Español de cuatro universidades taiwanesas son los que nos van a proporcionar la recogida de datos, incluyendo el siguiente número de estudiantes: 10 de la Universidad católica de Fu Jen, 59 de la Universidad de Tamkang, 88 de la Universidad del Providence, y 43 del Colegio universitario de los idiomas de Wenzao Ursuline del año 2005. Además, 89 de la Universidad de Tamkang, 144 de la Universidad del providence, y 192 del Colegio universitario de Wenzao Ursuline de 2006 son también objeto de nuestro análisis. (2) También recogemos los datos de los estudiantes españoles de los departamentos de idiomas extranjeros que constituyen el segundo grupo, los departamentos de lenguas extranjeras y el departamento de la educación general, incluyendo el siguiente número de estudiantes: 59 de la Universidad Nacional de Taiwán, 42 de la Universidad Nacional de Chengchi, 26 de la Universidad Nacional de Chiayi, 84 de la Universidad Nacional de Cheng Kung, 9 de la Universidad Nacional de Kaohsiung, 18 estudiantes de la Uni-

[9] Agradecimientos especiales a la coloboración de los Profesores Victorian Tian, Rita He, Ailin Yen, Francisco Moreno, Lucía Luo, Delia Lin, Teresa Chen, Javier Lu, Laura Vela, Edelmira Mao, Cecilia Liu, Azucena Lin, Eugenio Borao, Sofía Yang, Carolina Lin, Camilo Wang y Emilia Chen.

versidad de Da-Yeh, y 21 de la Universidad de Toko del año 2005. Además, incluimos a 93 de la Universidad Nacional de Taiwán, a 36 de la Universidad Nacional de Chengchi, a 79 de la Universidad nacional de Cheng Kung, a 14 de la Universidad de NanHua, a 16 de la Universidad nacional de Chiayi, a 14 de la Universidad China de Ultramar de la Tecnología, y a 13 de la Universidad Nacional de Formosa del año 2006. En total, los trabajos de los estudiantes de estas universidades alcanzan la cifra de 1.058 redacciones y de 209.486 palabras.

3.1.2. CATE-CIC: El corpus de aprendices taiwaneses de español-contraste entre incorrección y corrección

A través de la introducción de la versión corregida de los errores revisados por los hablantes nativos del español y de la aplicación de la función KWIC (Keywords in Context) de la Concordancia podemos obtener las palabras clave o sus estructuras, en las cuales los investigadores o los estudiantes están interesados. Con la recuperación de las palabras clave, la pantalla del ordenador está dividida en dos partes, en una de ellas muestra la oración original con errores y la otra la oración corregida, como se muestra en la siguiente imagen.



Antes de añadir las funciones de contraste de errores y la corrección, los investigadores pueden solamente consultar el CETE para recuperar los resultados del uso de los estudiantes y preparar los materiales didácticos. En el caso de querer hacer análisis más profundos, deberán hacerlo los propios investigadores sin la ayuda de KWIC debido a sus limitaciones. Por otra parte, los estudiantes tampoco se beneficiarán mucho de la consulta del corpus original.

Después de ser añadida la herramienta de contraste del error y corrección, los investigadores no sólo recuperan los datos usados por los estudiantes del español, sino que también pueden compararlos con las formas corregidas, o sea, con el uso de los hablantes nativos del español. Simultáneamente, los estudiantes también podrán beneficiarse mediante las consul-

tas de las correcciones que se corresponden con las equivocaciones cometidas por ellos mismos o sus compañeros de clase.

En los primeros dos años de la construcción del CATE, cien redacciones han sido comparadas con las formas corregidas por un profesor español nativo y marcadas con etiquetas que indican: los tipos de errores y las partes de la oración en las que se han cometido los mismos.

Señalando y anotando los errores de los estudiantes con etiquetas, ponemos al día las características técnicas del CATE y proporcionamos funciones más avanzadas para la búsqueda en el corpus y el análisis de los errores por parte de los investigadores y de los alumnos. En conclusión, el número de personas que pueden beneficiarse se ha ampliado notablemente.

3.2. Las funciones técnicas

3.2.1. Versión del web-site de CATE 2005-6

El CATE se divide principalmente en cuatro subsistemas: (1) interfaz de la recopilación, (2) creación del índice, (3) pregunta y (4) sistema de entrada. El lenguaje de programación para la recopilación del corpus es Perl y MySQL para la base de datos. Para que este recurso sea más manejable por parte de los usuarios se ha decidido que la interfaz posea la apariencia y la estructura de una página web.

3.2.1.1. Interfaz de la colección

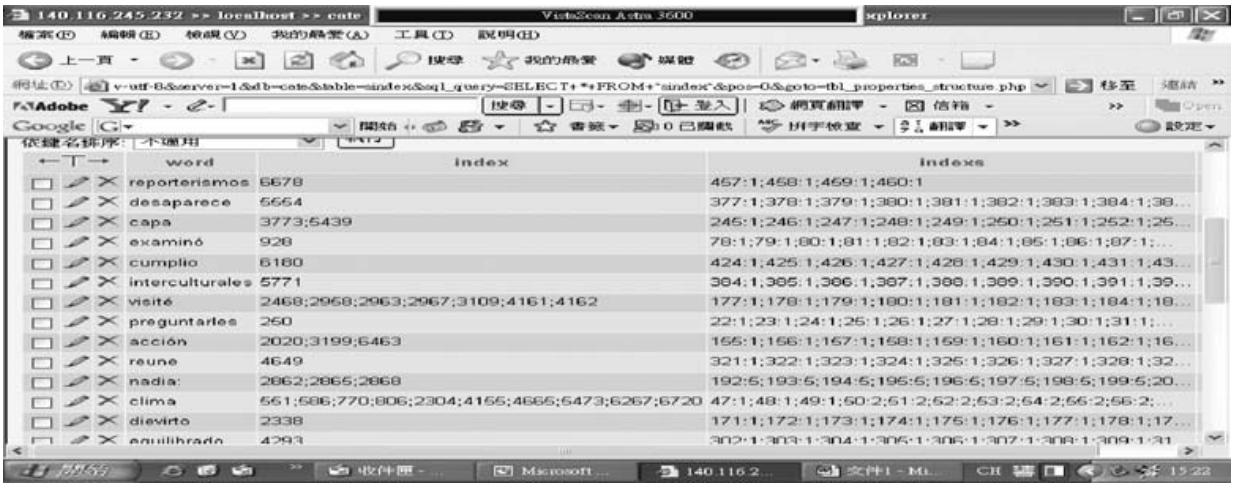
Debido a la variedad de autores y de fuentes de las que se ha surtido este trabajo de investigación, esperamos que al transcribir las redacciones, todos los archivos respeten las siguientes normas: (1) el código de la unidad y (2) las formas de la unidad.

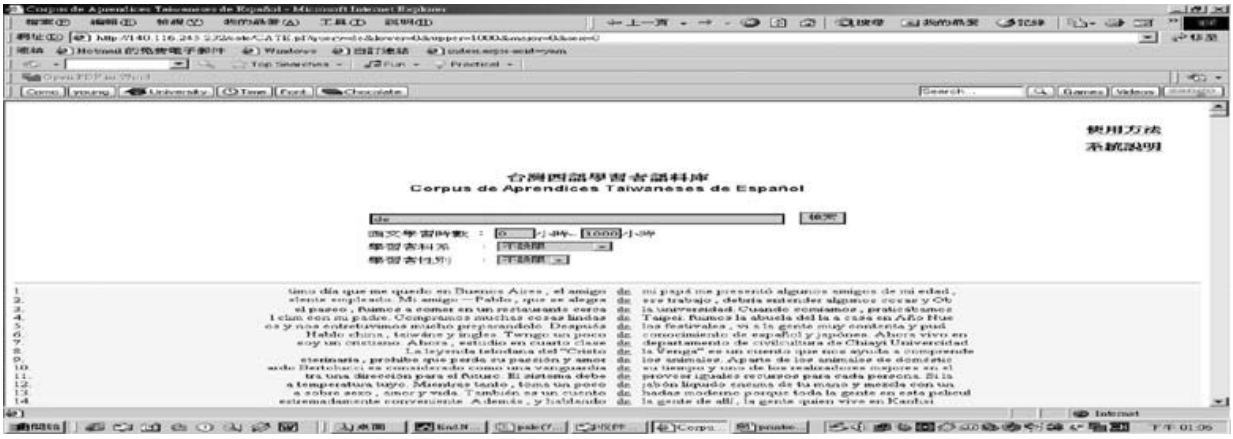
(1) Código de la unidad

Como los usuarios pueden utilizar diferentes códigos en sus redacciones, se debe utilizar el mismo código durante el proceso de recopilación, de lo contrario se podría producir una gran confusión en el índice. Por ejemplo, en el caso de asignar el código para la palabra con acento *sólo*, únicamente puede ser UTF-8 o ISO, porque en el caso de no usar el mismo código posteriormente, el programa distinguirá la misma palabra *sólo* como dos palabras diferentes. Así que constituimos un código de la unidad mediante la función de la codificación del web browser. De esta manera, los usuarios copian las redacciones al CGI, y el browser las transforma en el código que necesitamos para evitar el inconveniente de escribir el programa de transformación. Actualmente, el sistema utiliza UTF-8 para transcribir y consultar.

(2) Formas de la unidad

Como consecuencia del empleo de diferentes conceptos para efectuar la consulta (por ejemplo, se pueden consultar temas, autores...), surge la necesidad de dotar a los datos de unas determinadas formas que aporten unidad y que faciliten el método de consulta. Para dotar a este trabajo de una forma de unidad resulta conveniente crear la base de datos basándonos en las características de la construcción.





3.2.1.2. Creación del índice

El índice es uno de los elementos más importantes que constituyen el sistema de consulta. La función del índice es la de recordar la posición en la que podemos localizar cada palabra. Usando el sistema de consulta, podemos encontrar rápidamente las redacciones que contienen preguntas. Actualmente, guardamos el índice en la base de datos de MySQL, como se muestra en la siguiente imagen.

Cabe señalar que existen muchas maneras de guardar un índice, pero nosotros hemos considerado que el que más ayuda a nuestro propósito es el índice invertido. Un ejemplo del índice invertido sería el siguiente: suponemos que hay dos redacciones: la D1 y la D2, los datos estarían registrados en el índice invertido.

3.2.1.3. Interfaz de consulta

El sistema de la interfaz permite a los usuarios realizar la consulta para encontrar un determinado recurso. Además de la consulta básica, también existen otras opciones específicas de búsqueda, por ejemplo: horas de aprendizaje, los departamentos de los autores, temas, etc., como se señala en la siguiente imagen.

3.2.1.4. Sistema de introducción

Para evitar la posibilidad de que los usuarios puedan dañar el sistema, debe haber un mecanismo de identificación con el objetivo de proteger el sistema. Actualmente, se permite la entrada en el sistema a los usuarios que estén debidamente autorizados.

3.2.2. Versión de no-sitio-web

Con respecto a la versión de no-web-site, vamos a demostrar cómo podemos confeccionar un corpus y analizar datos eficientemente, con la menor ayuda posible de ingenieros profesionales en la lingüística computacional. Es decir, nos beneficiamos de las ventajas de las tres funciones principales proporcionadas por la herramienta del análisis léxico, la cuarta versión de WordSmith (programada por Mike Scott), a saber, WordList, Concord y KeyWords. La mayor dificultad con la que nos hemos topado es la de establecer un sistema de la codificación.

Usando las herramientas disponibles para analizar los datos guardados en los archivos con el formato del texto puro (archivo.txt), necesitamos prestar la atención al sistema de codificación. Después de investigar varios corpus creados y software aplicados, concluimos: por un lado, que el problema de la codificación se ha producido porque los sistemas de codificación mencionados interaccionan y entran en conflicto con los sistemas de los códigos que utiliza Microsoft. Mencionamos esta dificultad porque nunca habían sido contempladas en ningún manual de instrucción. Mientras que, por otro lado, las herramientas se pueden utilizar para estudiar datos españoles sin ningún problema bajo la versión inglesa o la española de Microsoft Office. Teniendo en cuenta nuestras experiencias previas con WordSmith, tenemos la posibilidad de aplicar el sistema de codificación del ISO para reconocer las letras especiales en español, á, é, í, ó, ú, ñ, ¡, aplicándolo al sistema chino de Windows.

4. ESTUDIO DE LAS ETAPAS DEL DESARROLLO

4.1. Propósito y cuestiones que se plantean en este estudio

El propósito del presente estudio consiste en desarrollar un orden jerárquico de aprendizaje a través del análisis de los errores cometidos por los estudiantes en lo referente a los usos de las palabras combinatorias. Partiendo de esta base, la pregunta fundamental que suscita el presente estudio es la siguiente: ¿cuáles son las categorías de las palabras combinatorias en las cuales cometen más errores los estudiantes y cuáles son las diferencias entre distintas etapas del aprendizaje?.

4.2. Estudiantes: CATE-CIC

En primer lugar, dividimos a los estudiantes en dos niveles: nivel I y nivel II, que se corresponden con los que llevan 128 horas estudiando el español y los que han aprendido el español durante 256 horas respectivamente. La manera de contar las horas de aprendizaje equivale a: 3 horas/semana x 16 semanas/semestre x 2 semestres/año x N año. Del nivel I, recogimos 37 redacciones, 530 oraciones y 10.432 palabras en total (6.193 palabras distintas); mientras que en el nivel II, sólo registramos 18 redacciones, 240 oraciones y 8.282 palabras (5.183 distintas palabras).

Los temas de las redacciones del nivel I son diferentes, por ejemplo: “El fin de semana”, “Un día en la universidad”, “Los días especiales de la pasada semana”, “El fin de semana pasado”, “La mirada nueva de nuestra habitación”, “Las cosas importantes para mí”, “Un paseo en bicicleta por la tarde”, “El verano pasado”, “Una tarde en la universidad”, “La fiesta para 2005”, “El día de final de año 2004”, ” Las vacaciones del verano pasado”, etc.

Por otro lado, los temas del nivel II incluyen: “Una comparición de las familias en China y en Japón”, “Un viaje solo”, “Mi experiencia de la educación en Taiwán”, “La gente de Taipei”, “Conflicto y coordinación del Romanticismo y la Realidad”, “Querer es Poder”, etc.

4.3. Metodología

Para contestar a las preguntas que se plantean en este estudio, analizaremos los datos extraidos del CETE-CIC anotándolos con etiquetas de errores y POS (las partes de la oración) con el objetivo de facilitar la presente investigación con la finalidad fundamental de aprovechar las herramientas de corpus existentes.

En primer lugar, clasificamos los errores cometidos por los estudiantes en tres grupos: el gramatical, el léxico y el ortográfico. Los errores gramaticales consisten en la concordancia, el pronombre reflexivo, el verbo (forma, tiempo, modo) entre otros. Los errores léxicos incluyen el vocabulario, la construcción, el sentido, el fragmento. Y por último, los errores ortográficos están compuestos por la puntuación, el acento, etc.

A continuación, las partes de la oración donde se cometen los errores se dividen en ocho categorías principales: el sustantivo, el adjetivo, el pronombre, el verbo, el adverbio, la preposición, la conjunción y la interjección.

Según los resultados obtenidos, podemos establecer los siguientes órdenes jerárquicos en función de la cantidad de errores cometidos. Para el nivel I, el orden desde lo más difícil a los más fácil es: Det-N (89), N-V (82), Pron-V (66), Adj-N (48), V-Adj (44), N-Adj (36), V-Adv (33), V-N (12), V-Conj (12),V-Prep-V (8), N-Prep-N (5).

Por otro lado, para el nivel II, el orden de dificultad es: Det-N (113), N-V (87), Adj- N (69), V-Adj (61), Pron-V (56), N- Adj (37),V-N (24), V-Adv (22), V-Conj (20), N-Prep-N (13), V-Prep-V (6).

Si comparamos los dos órdenes, nos percatamos de que son bastante similares, excepto algunos aspectos particulares (por ejemplo, el Pron-V). Por lo tanto, concluimos que la diferencia de cantidad de horas de estudio de español entre los dos niveles de estudiantes (entre los cuales hay una diferencia de 128 horas en el aprendizaje de la lengua española) no afecta mucho teniendo en cuenta los resultados de los usos, por parte de los dos niveles, de las palabras combinatorias tanto en lo referente a las colocaciones como en lo que se refiere a las coligaciones. Además, según la jerarquía de dificultad o de errores (Det-N, N-V, Adj-N, V-Adj, N-Adj, V- Adv/V-N, V-Conj, V-Prep-V/N-Prep-N), proponemos el siguiente órden de aprendizaje que iría desde una etapa más temprana a una más tardía (al revés de la de errores): V-Prep-V/N-Prep-N, V-Conj, V-Adv/V-N, N-Adj, V-Adj, Adj-N, N-V, Det-N.

A continuación, observaremos la relación que se establece entre las categorías de las palabras combinatorias y los tipos de errores. Según los resultados, nos dimos cuenta de que para los estudiantes del nivel I, los tipos de errores con mayor frecuencia varían entre el tipo léxico (5 de las 11 combinaciones listadas anteriormente), el gramatical (3 combinaciones) y el ortográfico (2 combinaciones). Con respecto a las combinaciones del tipo léxico, los errores tienden a ser de la clase "vocabulario".

Por otra parte, en cuanto al nivel II, observamos que la mayoría de los tipos de errores de mayor frecuencia pertenecen al tipo léxico (10 de 11 combinaciones), exceptuando una combinación de tipo gramatical. Además, dentro de las combinaciones del tipo léxico, los errores más frecuentes son de la clase "sentido".

Si cotejamos las diferencias entre los dos niveles, la dicción es el tipo de las palabras combinatorias en el que cometen errores con más frecuencia ambos niveles. Concentrándonos en el tipo léxico, notamos la diferencia entre los dos niveles: los estudiantes del nivel I tienden a cometer errores de la clase "vocabulario" mientras que los del nivel II, por los de la clase "sentido". En resumen, en las etapas de aprendizaje, existe un orden de facilidad a dificultad que va desde el tipo ortorgráfico hasta el gramatical y el léxico. Además, en cuanto al tipo léxico, la dificultad del sentido semántico aparece en una etapa más tardía que el vocabulario.

5. CONCLUSIÓN

Tomando como base el Corpus de Aprendices Taiwaneses de Español-Contraste entre Incorrección y Corrección, creado y anotado con etiquetas de errores y POS por Lu (2005-2006), y sirviéndonos de los instrumentos analíticos específicos de corpus, hemos pretendido estudiar el desarrollo léxico de los estudiantes que aprenden el español como segunda lengua extranjera.

Nos hemos centrado en el cotejo de las palabras combinatorias empleadas en distintos niveles de aprendizaje y propuesto un continum en lo que se refiere al grado de facilidad del proceso de aprendizaje, que sería el siguiente: V-Prep-V/N-Prep-N, V-Conj, V-Adv/V-N, N-Adj, V-Adj, Adj-N, N-V, Det-N. Además, no sólo nos percatamos de que en las etapas de aprendizaje, existe un orden de facilidad a dificultad que va desde el tipo otorgráfico hasta el gramatical y el léxico; sino que también especificamos que la dificultad del sentido semántico aparece en una etapa más tardía que el vocabulario con respecto al tipo léxico de las palabras combinatorias. Finalmente, queremos expresar nuestro deseo de que los resultados del presente trabajo faciliten tanto la comprensión del proceso de adquisición del español como segunda lengua como el diseño didáctico derivado de ésta.

REFERENCIAS BIBLIOGRÁFICAS

Aijmer, K., & A. B. Stenstrom (2004). *Discourse Patterns in Spoken and Written Corpora.* Amsterdam: J. Benjamins.

Al-Btoosh, M.A. (2005). “Interlanguage Lexicology of Arab Students of English: A Computer Learner Corpus-Based Approach”. *Dissertation Abstracts International, A: The Humanities and Social Sciences*, 66-1: 161-A-162-A.

Alexopoulou, A. (2005). “An Approach to the Treatment of Errors in Spanish as a Foreign Language Classes from the Perspective of Error Analysis”. *Estudios de Linguistica Aplicada,* 23: 101-125.

Chuang, Y. (1996). *Corpus Analysis of the Vocabulary in the Junior and Senior High School Students' English Textbooks and Writings in Taiwan*. Taipei: Crane.

Cobb, T. (2003). “Analyzing Late Interlanguage with Learner Corpora: Quebec Replications of Three European Studies”. *The Canadian Modern Language Review*, 59: 393-423.

Flowerdew, J. (2006). Use of Signalling Nouns in a Learner Corpus. *International Journal of Corpus Linguistics*, 11-3: 209-226.

Gonzalez-Espresati, C. (2002). Error Analysis and Interlanguage of Brazilian Learners of Spanish and Spanish Learners of Portuguese. *Hermeneu*, 4: 237-239.

Granger, S. (1998). *Learner English on Computer*. London, New York: Longman.

Huang, L.Y. (2000). “Corpora and Second Language Teaching and Learning”. Paper presented at the *QianXiNian Linguistics Conference*, National Cheng Chi University.

Kitsnki, M. (2006). “Language Corpora and Foreign Language Teaching”. *Eesti Rakenduslingvistika Uhingu Aastaraamat*, 2: 93-107.

Liang, M. (2006). “POS Tagging Reliability on EFL Learners' Written Data”. *Foreign Language Teaching and Research*, 38: 279-286.

Lisi, C.D. (2001). “English L2 Interlanguage Writing Development: Some Similarities and Differences among Spanish L1 Adolescent Learners”. *Dissertation Abstracts International, A: The Humanities and Social Sciences*, 62-4: 429-A.

Mönnink, I.D.; C. Mair, & M. Hundt (2000). “Parsing a Learner Corpus?”. In *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi: 81-90.

Neff, J.; F. Ballesteros; E. Dafouz; F. Martínez; J. P. Rica & M. Díez (2004). “Formulating Writer Stance: A Contrastive Study of EFL Learner Corpora”. In *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi: 73-89.

Neff, J.; E. Dafouz; H. Herrera; F. Martínez & J. P. Rica (2003). Contrasting Learner Corpora: The Use of Modal and Reporting Verbs in the Expression of Writer Stance. In Granger, S, & Petch-Tyson, S. eds. E*xtending the Scope of Corpus-based Research: New Applications, New Challenges*: 211-230. Amsterdam: Editions Rodopi B.V.

Scott, M. (1996). Oxford WordSmith Tools. Version 4. Oxford: Oxford University Press.

Tseng, Y. C. & H.C. Liou (2006). “The Effects of Online Conjunction Materials on College EFL Students' Writing”. *System,* 34: 270-283.

Wang, S.P. (1999).“Integration of Corpus-Based Approach into an EAP Class”. *Second Pan-Asia Conference-An Int. Forum*. Seoul, Korea.

http://140.116.245.232/japan_corpus

http://chss.montclair.edu/linguistics/MELD/

http://leo.meikai.ac.jp/~tono/lcorpuslist.html

http://mail.tku.edu.tw/dwible/index.htm

http://nora.hd.uib.no/corpora/1997-1/0262.html

http://www.uib.no/mailman/public/corpora-archive/2006-June/002850.html