LA CORRECCIÓN AUTOMÁTICA DE LA PRONUNCIACIÓN (EPAO). DISEÑO DE UN CORPUS ORAL DE MUESTRAS DE LENGUA DE ESTUDIANTES JAPONESES

Mario Carranza
Universidad de Estudios Extranjeros de Tokio

RESUMEN: Los últimos avances en el área de las tecnologías del habla junto con el creciente interés por el desarrollo de sistemas de enseñanza de la pronunciación asistidas por ordenador (EPAO) han propiciado la aplicación de estas tecnologías en la creación de nuevas herramientas y técnicas que hacen uso del reconocimiento automático del habla para la detención de errores de pronunciación en habla no nativa. En el presente artículo se abordarán aquellos aspectos que conciernen al diseño, la transcripción y el etiquetado de un corpus oral no nativo de producciones orales de estudiantes japoneses de español, de manera que puedan ser procesadas y utilizadas como corpus de entrenamiento de sistemas de reconocimiento del habla con aplicación en el desarrollo de herramientas de enseñanza de la pronunciación asistida por ordenador en el ámbito del español como lengua extranjera.

Palabras clave: enseñanza de la pronunciación asistida por ordenador, reconocimiento automático del habla, corpus oral no nativo, adquisición de la pronunciación del español LE, transcripción, anotación, etiquetado, japonés

Introducción

La enseñanza de lenguas asistida por ordenador (ELAO) se ha desarrollado notablemente en los últimos años, por lo que actualmente existe un amplio número de cursos en formato digital o en línea disponibles en el mercado, y varias instituciones académicas ofrecen parte de su docencia en la Red. No obstante, la enseñanza de lenguas asistida por ordenador se encuentra todavía en sus etapas iniciales y es necesario ahondar en aquellos aspectos que todavía no están resueltos de manera satisfactoria, como la enseñanza y la evaluación de la expresión oral. Una posible solución reside en utilizar la tecnología del reconocimiento automático del habla para desarrollar actividades que permitan la evaluación y la corrección de la pronunciación del estudiante. Sin embargo, el uso del reconocimiento automático del habla aplicado a locutores no nativos supone una serie de dificultades, puesto que los sistemas de reconocimiento se han entrenado únicamente con corpus orales nativos y las divergencias entre los dos tipos de señal son lo suficientemente amplias como para provocar un descenso drástico en el porcentaje de aciertos del reconocedor. Por lo tanto, para la aplicación de estas tecnologías en la enseñanza del español asistida por ordenador, se requiere que las herra-

mientas de reconocimiento se entrenen con corpus debidamente transcritos y etiquetados que reflejen las características de los usuarios, en este caso, hablantes no nativos con diversos niveles de dominio de la lengua.

La adaptación del reconocimiento automático del habla al desarrollo de herramientas de corrección automática de la pronunciación permite diseñar cursos en línea de español LE más específicos, orientados según la L1 del estudiante. El sistema es capaz de predecir los errores más comunes, así como de evaluar la pronunciación y proporcionar un feedback significativo que permita a los estudiantes entender la causa de sus errores de pronunciación y la manera de solucionarlos.

Los corpus orales de aprendices o corpus orales no nativos

Según la definición de Sinclair (1996), los corpus de aprendices se definen como «electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardised and homogeneous way and documented as to their origin and provenance.» Por lo tanto, se trata de recopilaciones de materiales reales de aprendices, compilados con un objetivo relacionado con la adquisición y enseñanza de L2/LE. Para cumplir este objetivo, los datos que forman el corpus deben ser codificados y documentados basándose en un estilo estandarizado, permitiendo así la compatibilidad del corpus y asegurando su reutilización para futuras aplicaciones. La recopilación de un corpus es un trabajo complejo, que requiere una importante inversión económica y una dedicación plena, por lo que es absolutamente necesario considerar en un primer lugar cuáles son los objetivos para la creación de ese corpus y diseñarlo a partir de unos criterios adecuados para que se adapte a esos objetivos. Un corpus debidamente transcrito y anotado no solo es válido para el objetivo original a partir del cual se diseñó sino que representa una valiosa fuente de información para cualquier otra investigación con un enfoque similar; incluso, puede ser utilizado en el ámbito del aula como herramienta de consulta, lo que permite obtener de manera rápida y sencilla datos reales de las producciones, tanto de hablantes nativos, como de aprendices de una lengua extranjera (Sinclair, 2004).

En el caso que nos ocupa, la investigación con corpus de aprendices es una disciplina relativamente joven, que se empezó a desarrollar en 1980, vinculando por primera vez el campo de la lingüística de corpus con el campo de la investigación en la adquisición y enseñanza de lenguas extranjeras. Una de las ventajas de la investigación con corpus es que permite dotar cualquier investigación de una base empírica, lo que, en ocasiones, sirve para corroborar una teoría, hasta ese momento puramente intuitiva, o para plantear nuevos modelos teóricos de adquisición de las lenguas extranjeras. Asimismo, los corpus de aprendices han sido utilizados en gran medida para el diseño curricular de cursos en línea, la preparación de materiales o la redacción de diccionarios y gramáticas orientados a los estudiantes extranjeros (Meunier, 2002) y se han adaptado como corpus de entrenamiento en el desarrollo de tecnologías lingüísticas debido a su capacidad predictiva (Granger, 2002). La investigación con corpus de aprendices o no-nativos posee un alto grado de interdisciplinaridad, puesto que su estudio es objeto de tres grandes disciplinas: los modelos de adquisición de L2 y LE, la enseñanza de las lenguas extranjeras y, por último, el procesamiento del lenguaje natural (véase figura 1). La progresiva adaptación de las tecnologías del habla a la ense-

ñanza de lenguas extranjeras, en especial, los primeros intentos de utilizar el reconocimiento automático del habla como herramienta para la enseñanza y corrección de la pronunciación, propiciaron la necesidad de entrenar al reconocedor con corpus orales no nativos, de forma tal que el sistema de reconocimiento sea capaz de distinguir el habla nativa del habla no nativa y pueda detectar aquellos errores de pronunciación más frecuentes en el habla de un extranjero.

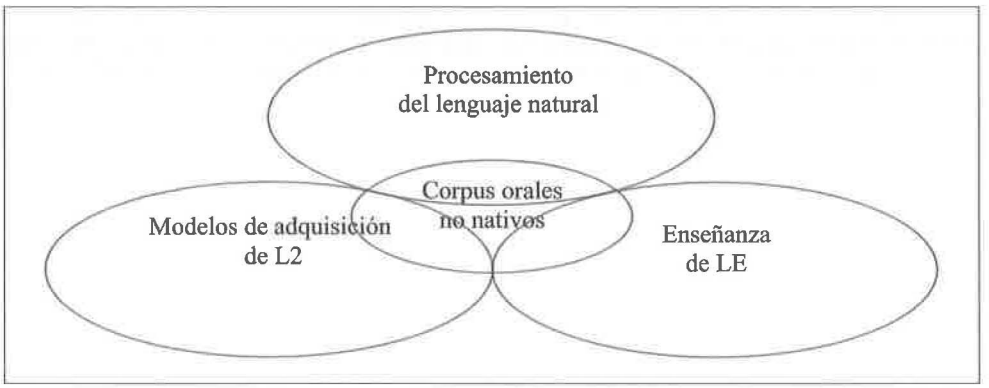

Figura 1.: *Disciplinas relacionadas con el estudio de los corpus orales no nativos*

Actualmente, existen varios proyectos en funcionamiento que se ocupan de la recopilación de corpus orales no nativos, tanto en el campo de la metodología de enseñanza de la pronunciación de lenguas extranjeras como en el campo de las tecnologías del habla. En primer lugar, es necesario mencionar algunos proyectos destacados que se han llevado a cabo en Europa como, por ejemplo, el corpus ISLE (Interactive Spoken Language Education) de inglés como lengua extranjera, cuyo objetivo es la recopilación de datos orales de inglés LE de estudiantes de L1 alemán y de L1 italiano con el propósito de desarrollar herramientas para la enseñanza de la pronunciación (Menzel et al., 2000). Los datos de este corpus, realizado por la Universidad de Leeds, la Universidad de Milán y la Universidad de Hamburgo y distribuido por la Agencia Europea de Distribución de Recursos Lingüísticos (ELDA), se han utilizado para evaluar el nivel de precisión en tareas de reconocimiento de habla, entrenar y evaluar reglas de localización de errores fonéticos y evaluar procedimientos de detención del acento léxico.

Las investigaciones sobre corpus orales no nativos en lenguas que no sean inglés son todavía escasas. Disponemos de pocos corpus orales y los que existen no son fácilmente accesibles o son de un tamaño muy limitado. En el campo del francés como lengua extranjera conviene destacar el proyecto IPFC (Interphonologie du Français Contemporain) , desarrollado por varias universidades europeas, que está orientado a la recopilación de datos orales de francés no nativo, como lengua extranjera o como segunda lengua. El corpus recopilado en este proyecto permite el análisis lingüístico a diferentes niveles de la interlengua de estudiantes extranjeros de francés en función de

la L1 del estudiante. Asimismo, se ofrece como una herramienta que permite dotar de una base empírica a cualquier teoría metodológica o a los modelos de adquisición del sistema fonético-fonológico del francés como lengua extranjera (Racine et al., 2011).

El francés LE y el inglés LE también son las lenguas analizadas en el corpus COREIL, cuyo objetivo es el estudio de la adquisición del aspecto prosódico de una lengua extranjera. El protocolo de recolección de datos orales, diseñado para el proyecto permite adaptarse a diferentes tipos de locutores según la edad, el grado de dominio de la L2; así como su L1 y al tipo de tarea. A pesar de que el proyecto todavía no está terminado, el concepto de un protocolo modular, que permita poder comparar diversos parámetros entre sí y pueda ser adaptado a cualquier tipo de corpus oral –incluyendo corpus nativos–, resulta una idea innovadora y quizás se consolide, en un futuro, como un estándar en la anotación de corpus orales no nativos (Delais-Roussarie y Yoo, 2011).

Por último, cabe apuntar el proyecto EURONOUNCE, cuyo diseño se basa en la creación de un sistema inteligente de enseñanza de lenguas extranjeras con un módulo de corrección automática de la pronunciación que integra herramientas de reconocimiento automático del habla y detención automática de errores. Este proyecto se centra en las lenguas polaco, sueco, ruso y eslovaco para hablantes de L1 alemán y alemán LE para hablantes de las lenguas antes mencionadas (Cylwik et al., 2009).

Además de los proyectos presentados en este apartado, existen otros corpus orales de aprendices desarrollados para el italiano, alemán, neerlandés, coreano y noruego. Para una mayor información recomendamos consultar la lista de corpus orales de aprendices publicada en la página web de la Universidad Católica de Lovaina[1], que se actualiza con frecuencia.

Proyectos de corpus orales de aprendices de español LE desarrollados en el ámbito europeo

A pesar de que el español es una de las lenguas más estudiadas en todo el mundo, el interés por la recopilación de corpus orales de aprendices ha sido relativamente escaso. Por ahora, solo disponemos de dos proyectos centrados en la recopilación de corpus orales de aprendices de español como lengua extranjera: el corpus SPLLOC y el corpus LUIS. Existen, no obstante, otros proyectos de corpus multilingües de aprendizaje de lenguas europeas, como el corpus C-ORAL-ROM o el Multilingual Learner Corpus (MLC), con los que se pueden obtener datos orales de aprendices de español LE. A continuación, comentaremos brevemente los objetivos y el desarrollo de estos proyectos.

Corpus LUIS-L2/LE(Language Usage In Spanish L2)

Proyecto desarrollado en colaboración entre la Universidad Pompeu Fabra y la Escuela Oficial de Idiomas de Barcelona-Drassanes, que consiste en la elaboración de un corpus oral de producciones de estudiantes de español como LE en tareas semi-

[1] Disponible en [http://www.uclouvain.be/en-cecl-lcWorld.html].

espontáneas, en una perspectiva longitudinal. El corpus está transcrito ortográficamente y fonéticamente y cuenta con una anotación de errores más frecuentes. La transcripción ortográfica se ha realizado con el estándar CHILDES[2], un sistema de codificación y transcripción del habla infantil empleado internacionalmente en los estudios de adquisición del componente fónico de la L1. El corpus se puede consultar en Internet desde la base de datos Talkbank[3], aunque se ha publicado también en formato libro, acompañado de un CD con una selección de los datos orales utilizados en el corpus (Díaz, 2007). Actualmente, el corpus cuenta con datos de hablantes de alemán, sueco, islandés, coreano y chino y próximamente se ampliará con hablantes de otras lenguas.

Corpus SPLLOC (Spanish Learner Language Oral Corpora) de aprendices nativos de inglés

Se trata de un proyecto desarrollado por la Universidad de York, la Universidad de Southampton y la Universidad de Newcastle y financiado por el Economic & Social Research Council, con el objetivo de recopilar un corpus oral no nativo para investigar sobre la adquisición de aspectos morfosintácticos del español como lengua extranjera por hablantes de inglés L1[4]. El corpus ofrece los datos en distintos formatos de audio, así como la transcripción ortográfica, para la que se utiliza el estándar CHILDES, al igual que el corpus LUIS. Puede consultarse en línea, directamente desde la página web del proyecto[5], como a través de la página web de las bases de datos Talkbank.

Corpus oral multilingüe C-ORAL-ROM (Integrated Reference Corpora for Spoken Romance Languages)

Este proyecto de creación de un corpus oral multilingüe de las lenguas romances, con muestras procedentes de hablantes nativos de portugués, español, francés e italiano, pero con una explotación didáctica como material orientado a la comprensión oral en LE. La parte correspondiente al español fue desarrollada por un equipo de investigadores de la Universidad Autónoma de Madrid dirigido por el profesor Antonio Moreno, y se ha publicado en formato digital como material complementario de un libro didáctico sobre enseñanza del español oral (Campillo et al., 2010). Los datos no se encuentran disponibles en Internet pero puede obtenerse más información acerca del proyecto en su página web[6].

---

[2] El sistema de transcripción CHILDES es usado generalmente en los corpus de lenguaje infantil, aunque se ha utilizado también para la transcripción de corpus orales no nativos y corpus orales clínicos de hablantes afásicos. Se puede consultar toda la información referida a este sistema de transcripción en la siguiente página web [http://childes.psy.cmu.edu/].

[3] Disponible en [http://talkbank.org/SLABank/].

[4] Los resultados de este análisis pueden consultarse en Mitchell et al. (2008).

[5] Disponible en [http://www.splloc.soton.ac.uk/search.php].

[6] [http://lablita.dit.unifi.it/coralrom/].

EL CORPUS ORAL DE APRENDICES JAPONESES DE ESPAÑOL LE «TUFS»

El origen de la creación de un corpus oral de estudiantes japoneses de español como lengua extranjera se remonta a una base de datos digital de producciones orales de los estudiantes japoneses de primer y segundo curso recopilada por el Departamento de Español de la Universidad de Estudios Extranjeros de Tokio (TUFS). Estos datos están siendo obtenidos a lo largo de los cursos 2010-2011 y 2011-2012. El objetivo principal consistía originalmente en ofrecer estos datos a la comunidad investigadora a través de un soporte digital; sin embargo, debido al gran tamaño de memoria necesaria para la grabación de estos archivos en un soporte físico, se optó por la solución de componer un corpus oral con las grabaciones debidamente transcritas y etiquetadas, que sirviera como complemento al Corpus CbLLE de producciones de hablantes nativos de español, ya desarrollado por esta misma universidad[7], y accesible desde Internet. Posteriormente, se decidió ampliar los niveles de transcripción, etiquetado y codificación de manera que el corpus, en un futuro, permita su explotación, no solo como una base de datos orales para cualquier estudio sobre la pronunciación del español por hablantes de japonés, sino también como un corpus de entrenamiento de un sistema de reconocimiento automático del habla, que permita la aplicación de estos sistemas en el desarrollo de herramientas de enseñanza y corrección de la pronunciación del español como lengua extranjera. Para ello, se han seleccionado 30 informantes, (15 hombres y 15 mujeres) de entre un total de 150 locutores y se ha procedido a realizar la transcripción ortográfica y fonética de tres tareas orales: conversación semi-espontánea, conversación espontánea y lectura. Las grabaciones han sido tomadas en cuatro momentos diferentes durante los dos años de estudio de español, por lo que para cada locutor se dispone de un total de doce grabaciones.

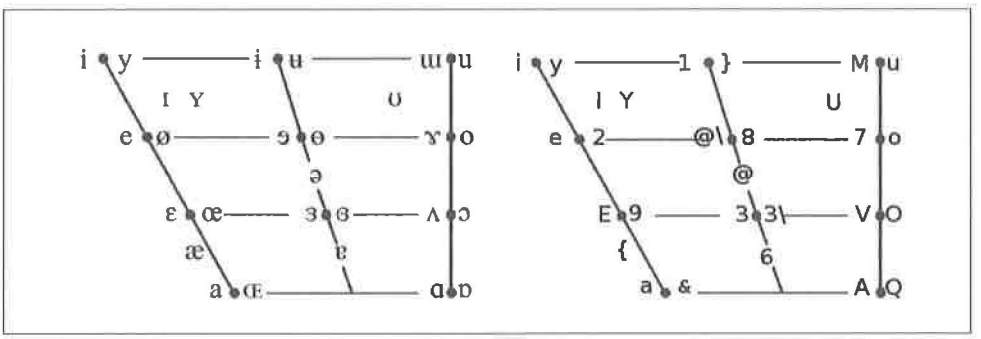

Figura 2: *Tabla de símbolos fonéticos para las vocales en AFI (izquierda) y X-SAMPA (derecha)*[8]

[7] Disponible en línea en [http://cblle.tufs.ac.jp/tag/es/search.php?menulang=en].

[8] Obtenido de la página web del Laboratorio de Fonética Experimental «Arturo Genre» de Turín, disponible en [http://www.lfsag.unito.it/ipa/index_en.html].

Transcripción ortográfica y fonética

La transcripción fonética de un corpus no nativo es una tarea compleja, que requiere, en primer lugar, establecer el inventario de unidades fonéticas que se va a utilizar para la transcripción. El habla no nativa o extranjera presenta un alto grado de variabilidad fonética, puesto que el estudiante percibe las categorías fonológicas de la lengua extranjera a través del sistema fonológico de su L1, que actúa como un filtro, por lo que el estudiante puede resultar «sordo» a ciertos sonidos, o matices, de la LE. En la producción oral, las categorías fonológicas de la L1 actúan a l manera de imanes, atrayendo a los sonidos de la LE que se articulen en una posición cercana; el resultado es la transferencia negativa del sistema fonológico de la L1 en la producción de los sonidos de la lengua extranjera. Por lo tanto, a la hora de decidir qué categorías deben constituir el inventario fonético para la transcripción del corpus no basta con considerar los sistemas fonológicos de la L1 y la L2, ya que existe un rango de variación muy amplio entre las realizaciones de los estudiantes y las realizaciones nativas. Por otro lado, el sistema fonológico de la interlengua posee una variación intra-locutor y es dinámico; es decir, que los errores no son constantes y cambian a lo largo del tiempo de instrucción. Esto se refleja en un alto grado de variación no solo entre aprendices con el mismo nivel de instrucción sino también en el rendimiento individual de cada aprendiz. Cabe considerar, por otro lado, la necesidad de utilizar un conjunto de símbolos para la transcripción fonética que puedan ser procesados por el sistema de reconocimiento (machine-readable), por lo que el inventario debe estar basado en caracteres alfanuméricos (esto descarta el uso de otro tipo de caracteres como los símbolos fonéticos de la Asociación Fonética Internacional, AFI), a la vez que debe ser capaz de representar una transcripción fonética estrecha, de manera que el inventario cuente con un número suficiente de símbolos para representar de la forma más fidedigna posible la realización del hablante. En el ámbito europeo se desarrolló un sistema alfanumérico para realizar transcripciones fonológicas de las lenguas europeas, conocido como SAMPA, cuya versión ampliada (X-SAMPA) cuenta con suficientes caracteres para representar los símbolos de la AFI, así como los diacríticos, que permiten una transcripción mucho más detallada (Wells, 1994). El inventario de SAMPA –propio de cada lengua[9]– se ha utilizado comúnmente para la transcripción de corpus orales nativos orientados al entrenamiento de sistemas de reconocimiento automático del habla, mientras que el inventario ampliado de X-SAMPA surgió de la necesidad de realizar una transcripción estrecha. Ambos sistemas de representación permiten la compatibilidad entre sí, lo que asegura la posibilidad de recuperar el corpus para futuras investigaciones (Llisterri, 1999). En la figura 2 se puede observar la relación entre algunos fonemas de la AFI, y su correspondencia con los símbolos de X-SAMPA. Para la transcripción fonética del corpus TUFS se decidió utilizar un sistema de codificación basado en el estándar SAMPA para el español (Wells, 1986), ampliado con algunos caracteres de X-SAMPA, de manera que permita anotar las características acústicas de los errores de pronunciación de los estudiantes.

[9] La adaptación de SAMPA al español puede consultarse en Wells (1986) disponible en [http://www.phon.ucl.ac.uk/home/sampa/spanish.htm].

Etiquetado y codificación de los errores

La mayor parte de los software orientados a la transcripción de corpus orales permiten codificar automáticamente una serie de informaciones como el locutor, el entrevistador, el año de estudio y el sexo del estudiante, y el tipo de tarea. Una vez que el corpus se haya transcrito ortográfica y fonéticamente, se compara con una transcripción fonética de referencia que contiene la realización canónica de un hablante nativo y que es generada automáticamente a partir de la transcripción ortográfica. A partir de esta comparación, se identifican las divergencias entre la realización del estudiante y la realización canónica del nativo y estas divergencias sirven de base para establecer los errores de pronunciación prototípicos de los estudiantes. Los tipos de errores encontrados en el habla de los estudiantes japoneses se han establecido considerando el nivel fonémico, a partir de tres categorías[10]:

- Errores debidos a la elisión de un fonema, en aquellos contextos en los que una secuencia de fonemas es difícil de pronunciar por el estudiante y la secuencia se simplifica eliminando uno o más fonemas.
- Errores debidos a la sustitución de un fonema por otro, en la realización de fonemas no existentes en la L1 o en aquellos fonemas que, a pesar de existir en las dos lenguas, se realizan articulatoriamente de manera diferente.
- Errores debidos a la inserción de un fonema nuevo, especialmente en las combinaciones de fonemas que no son permitidas en la L1, el estudiante tiende a incluir un fonema de apoyo a fin de reestructurar la secuencia de manera que resulte en una combinación posible en su L1.

Para la codificación de los errores se ha optado por utilizar un código numérico de siete cifras, donde la primera cifra hace referencia al tipo de error; la segunda y la tercera, al fonema implicado; la cuarta y la quinta cifra, al contexto fonético precedente y la quinta y sexta cifra al contexto fonético posterior. Por ejemplo, la realización de la palabra «suponemos» con una elisión de la primera vocal [u] se codificaría como 1301601, cifra que corresponde a una elisión [1] del fonema /u/ [30] entre una consonante fricativa palatal sorda /s/ [16] y una consonante oclusiva labial sorda /p/ [01][11]. A pesar de que todavía debemos probar este sistema de codificación en el corpus, que podrá ser ajustado o ampliado en un futuro, consideramos que el sistema de codificación adoptado permitirá relacionar de manera satisfactoria los tipos de errores y desviaciones respecto a la pronunciación canónica, que reflejan los datos de la señal acústica. Es decir, el ejemplo anterior muestra que este sistema de codificación de errores permite al reconocedor saber que en ese contexto existe un sonido [u] que se ha elidido, información que se perdería con una mera transcripción fonética y además nos proporciona información sobre los contextos de aparición más frecuentes de este error.

[10] En Carranza (en prensa) se puede encontrar una revisión bibliográfica sobre los estudios fonético-fonológicos contrastivos entre el español y el japonés y otros estudios de carácter metodológico sobre la enseñanza de la pronunciación española a hablantes de japonés. No obstante, el desarrollo del corpus de realizaciones orales de estudiantes japoneses de español nos permitirá verificar de manera empírica la presencia de estos errores en el habla de los estudiantes, así como obtener datos sobre su frecuencia de aparición y los contextos fonéticos que puedan propiciarlos.

[11] Este sistema es el utilizado en el proyecto IPFC, consúltese Detey (en prensa) para una explicación más detallada.

El proyecto de creación del corpus TUFS de realizaciones orales de aprendices japoneses de español está todavía en elaboración. Una vez se haya concluido el proyecto se dispondrán de todos los datos transcritos ortográfica y fonéticamente a través de una página web, desde la que se podrán descargar y ser utilizados para cualquier investigación enfocada a la adquisición de la pronunciación del español por hablantes de lengua japonesa.

Conclusiones

El creciente interés por la lingüística de corpus y el uso de corpus en el aula de lenguas extranjeras está vinculado directamente al desarrollo de la tecnología. Tradicionalmente, el desarrollo de un corpus suponía una inversión considerable en tiempo y esfuerzo por parte del investigador. Sin embargo, en la actualidad el progreso tecnológico ha hecho posible que el almacenamiento de grandes bases de datos pueda realizarse de manera sencilla en un ordenador personal. A su vez, existen numerosas herramientas disponibles en el mercado que permiten el análisis automático o semi-automático de los datos, lo que ha propiciado un avance en la investigación sobre corpus y su aplicación en la enseñanza de lenguas extranjeras. Los últimos avances tecnológicos cada vez se implantan en más aspectos de nuestra vida diaria; un ejemplo de ello es el uso del reconocimiento automático del habla, que se comenzó a utilizar principalmente para los sistemas de diálogo, y hoy en día esta tecnología se emplea en otros muchos campos: traducción automática, dictado automático y asistentes virtuales en teléfonos móviles. A pesar de que su aplicación a la corrección y enseñanza de la pronunciación de lenguas extranjeras todavía se encuentra en una etapa inicial, el entrenamiento de sistemas de reconocimiento mediante corpus orales no nativos permitiría un diseño de cursos en línea enfocado a los problemas individuales del aprendiz, y proporcionaría el feedback adecuado para la corrección paulatina de sus errores de pronunciación. El aumento de proyectos relacionados con la compilación de corpus orales no nativos para el entrenamiento de sistemas de reconocimiento automático del habla desarrollados en el ámbito europeo indica que en un futuro próximo contaremos con estas herramientas y podremos utilizarlas en las clases de ELE tanto presenciales como en línea. No obstante, el limitado interés que han suscitado los corpus orales de aprendices en la investigación sobre enseñanza y corrección de la pronunciación del español como lengua extranjera puede suponer un obstáculo para el desarrollo futuro de estas herramientas. Por el momento, contamos con dos corpus orales de español diseñados para la investigación sobre la adquisición del componente morfosintáctico; además, la falta de una transcripción fonética realizada con sistemas de codificación estandarizados impide adaptar estos corpus como sistema de entrenamiento de un reconocedor. Es necesario un mayor número de proyectos de corpus orales de aprendices de español transcritos es sistemas estandarizados que permitan su uso como corpus de entrenamiento de sistemas de reconocimiento de habla.

Los corpus de aprendices o corpus no nativos representan una herramienta robusta para el análisis y la descripción de la interlengua de los estudiantes. En primer lugar, un corpus anotado y etiquetado permite obtener la frecuencia de aparición de aquellos errores más característicos en función de diferentes parámetros, como el tipo de tarea, el nivel de dominio de la lengua o incluso el sexo, lo que nos

indicará qué errores son recurrentes a lo largo del aprendizaje del español y cuáles se suelen corregir con la instrucción. Por otro lado, dispondremos de los contextos de aparición característicos de cada error, esta información posee un valor notable para el diseño de programas y materiales de enseñanza y corrección de la pronunciación, puesto que permite saber qué tipo de tareas y contextos son los más adecuados para corregir los errores en la pronunciación del español por hablantes de una determinada lengua. Por último, el análisis estadístico de un corpus de aprendices puede apuntar tendencias globales en el desarrollo de la adquisición del componente fónico del español, lo que abrirá camino a futuras investigaciones que permitan describir con datos empíricos la adquisición del componente fónico del español como lengua extranjera.

Bibliografía

CAMPILLOS LLANOS, Ignacio, Paula GOZALO GÓMEZ, José María GUIRAO y Antonio MORENO SANDOVAL (2010): *Español oral en contexto 1. Textos de español oral,* Madrid: UAM Ediciones.

CARRANZA, Mario (en prensa): *Errores y dificultades específicas en la adquisición de la pronunciación del español LE por hablantes de japonés y propuestas de corrección.*

CYLWIK, Natalia, Agnieszka WAGNER, y Grazyna DEMENKO (2009): «The EURONOUNCE corpus of non-native Polish for ASR-based pronunciation tutoring system», *SLaTE 2009. ISCA workshop on speech and language technology in education,* Warwickshire: England. Disponible en <http://www.eee.bham.ac.uk/SLaTE2009/papers\SLaTE2009-30.pdf>

DELAIS-ROUSSARIE, Elizabeth y Hi-Yon YOO (2011): «Learner corpora and prosody: From the COREIL corpus to principles on data collection and corpus design», *Poznán Studies in Contemporary Linguistics*, 47(1), 26-39.

DETEY, Sylvain. (en prensa): *Coding an L2 phonological corpus: From perceptual assessment to non-native speech models – an illustration with French nasal vowels.*

DÍAZ RODRÍGUEZ, Lourdes. (2007): *Interlengua española – Estudio de casos,* Barcelona: Regael.

GRANGER, Sylviane, Joseph HUNG y Stephanie PETCH-TYSON (Eds.) (2002): *Computer learner corpora, second language acquisition and foreign language teaching,* Amsterdam: John Benjamins.

GRANGER, Sylviane (2002): «A bird's-eye view of learner corpus research», en Sylviane GRANGER, Joseph HUNG y Stephanie PETCH-TYSON (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching*, Amsterdam/Philadelphia: John Benjamins, 3-33.

KAWAGUCHI, Yuji, Susumu ZAIMA y Toshihiro TAKAGAKI (2006): *Spoken language corpus and linguistic informatics,* Amsterdam: John Benjamins.

LLISTERRI, Joaquim (1999): «Transcripción, etiquetado y codificación de corpus orales», en J. GÓMEZ GUINOVART, A. M. LORENZO SUÁREZ, J. PÉREZ GUERRA y A. ÁLVAREZ LUGRÍS (Eds.) *RESLA, Revista Española de Lingüística Aplicada*, Volumen monográfico *«Panorama de la investigación en lingüística informática»,* 55-82.

MENZEL, Wolfgang, Eric ATWELL, Patrizia BONAVENTURA, Daniel HERRON, Peter HOWARTH, Rachel MORTON y Clive SOUTER, C. (2000): «The ISLE corpus of non-native spoken English», en *Proceedings of the Second International Conference on Language Resources and Evaluation,* mayo de 2000, Atenas.

MEUNIER, Fanny (2002): «The pedagogical value of native and learner corpora in EFL grammar teaching», en Sylviane GRANGER, Jospeh HUNG y Stephanie PETCH-TYSON (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching*, Amsterdam: John Benjamins, 119-141.

MITCHELL, R., DOMÍNGUEZ, L., ARCHE, M. J., MYLES, F., y MARSDEN, E. (2008): «SPLLOC: A new database for Spanish second language acquisition research», *EUROSLA Yearbook*, 8, 287-304.

RACINE, Isabelle, Françoise ZAY, Sylvain DETEY, y Yuji KAWAGUCHI (2011): «De la transcription de copus à l'analyse interphonologique: enjeux méthodologiques en FLE», en G. COL y S. N. OSU (Eds.), *Transcrire, écrire, formaliser (1). Travaux Linguistiques du CerLiCO* 24, Rennes: PUR, 13-30.

SINCLAIR, John (1996): *EAGLES. Preliminary Recommendations on Corpus Typology*. Disponible en <www.ilc.cnr.it/EAGLES/pub/eagles/corpora/corpustyp.ps.gz>

SINCLAIR, John (Ed.) (2004): *How to use corpora in language teaching. Studies in Corpus Linguistics*, 11. Amsterdam: John Benjamins.

WELLS, John (1986): SAMPA for Spanish. Department of Phonetics and Linguistics, University College London. Disponible en <http://www.phon.ucl.ac.uk/home/sampa/spanish.htm>

WYNNE, Martin (Ed.). (2005): *Developing linguistic corpora: A guide for good practice,* Oxford: Oxbow Books.