Un corpus de blogs de aprendices japoneses de español

María del Pilar Valverde Ibañez
Universidad de Estudios Extranjeros de Kansai

1. Introducción

En la última década la creación de corpus de aprendices, primero para el inglés pero más tarde también para otras lenguas, ha ido en aumento. Desde la aparición del *International Corpus of Learner English* en 2009, los corpus de aprendices han abierto nuevas vías de investigación (Granger *et al*. 2005) en el campo de la lingüística de corpus, la adquisición de segundas lenguas, la enseñanza de lenguas extranjeras y el procesamiento del lenguaje natural.

Los corpus demuestran ser necesarios para la investigación en adquisición de lenguas extranjeras, ya que son una fuente de datos abundante, representativa y fiable (Mendikoetxea 2014; Alonso-Ramos 2016). También son de gran utilidad para la creación de materiales de enseñanza que presten atención a las áreas más problemáticas para los aprendices (como los materiales de Cambridge University Press, basados en el Cambridge Learner Corpus). Así mismo, el enriquecimiento de los textos con información lingüística y con la detección automática de errores gramaticales de los aprendices (Leacock *et al.* 2014) aumenta sus posibilidades de explotación.

1.1. Corpus de aprendices del español

Para el español, el número de corpus de aprendices es todavía escaso. Los proyectos más remarcables de corpus escritos son el corpus CEDEL2 (Lozano 2009) (750.000 palabras de aprendices cuya primera lengua es el inglés), el corpus CATE (Lu 2010) (300.000 palabras de aprendices hablantes de chino) y el corpus CAE (Rojo y Palacios 2016) (500.000 palabras de aprendices cuya primera lengua es el inglés, francés, portugués, árabe, chino o ruso). Hasta el momento no existe ningún corpus de libre acceso que contenga datos de aprendices de español cuya lengua materna sea el japonés.[1]

En el marco de un proyecto de investigación para la creación de un corpus de aprendices de español cuya lengua materna es el japonés,[2] con el objetivo de facilitar la investigación sobre la adquisición del español y la mejora de la enseñanza del español en Japón, en este artículo exploramos la posibilidad de tomar algunos textos disponibles en la web en forma de blogs para su inclusión en el corpus.

1. El corpus CORANE (Mancera *et al.* 2001) contiene algunos textos producidos por este tipo de aprendices –entre otros–, concretamente 58 aprendices cuya lengua materna es el japonés. Sin embargo, se trata de un corpus de acceso limitado y de una muestra que resulta insuficiente para llevar a cabo investigaciones más amplias.
2. Grant-in-Aid for Scientific Research (Start-up) de la Japan Society for the Promotion of Science (17H07270).

1.2. El uso de la web como corpus

La web constituye una fuente inagotable de datos sobre la lengua, como muestra la larga tradición de desarrollo de web corpus y de herramientas para su consulta cuenta. Sin embargo, esta fuente de datos ha recibido poca atención en la construcción de corpus de aprendices.[3] Los corpus de aprendices suelen estar compuestos por tareas de clase o exámenes, realizados en casa o en el centro de enseñanza, con o sin limitación de tiempo o de acceso a materiales de consulta, debido a que en este tipo de corpus es necesario controlar múltiples variables relacionadas con el aprendiz y con la tarea.

En esta investigación, compilamos un corpus formado textos publicados en 48 blogs, esto es, páginas web, generalmente personales, con una estructura cronológica, que se actualizan regularmente y que se suele dedicar a un tema concreto, y cuyos autores son aprendices de español hablantes de japonés como lengua materna.

Los blogs, a diferencia de los textos producidos típicamente en un contexto de enseñanza:

a. Son escritos por la propia iniciativa de los aprendices.

b. Dan libertad de elección de tema y estilo al autor, como ocurre en los diarios personales tradicionales. Por ello, los blogs son una buena fuente de información sobre las motivaciones de los aprendices para escribir, sus temas preferidos e incluso su personalidad (Gil *et al.* 2009).

c. Presentan rasgos tanto del registro escrito como del registro oral (Herring *et al*. 2005). Tienden a ser un registro informal (la lengua usada en ellos está menos sujeta a reglas que el registro académico, por ejemplo), pero más cuidado desde el punto de vista de la gramática y la estructura que otros géneros electrónicos como los *tweets* u otras publicaciones en redes sociales.

d. Son un monólogo y un diálogo al mismo tiempo (Efimova y Moor 2005). A diferencia de los autores de diarios personales tradicionales que generalmente se mantienen en privado, los autores de diarios personales en línea, son conscientes de que su textos serán públicos, es decir, de la presencia de potenciales lectores y sus posibles reacciones, por lo que reflexionan sobre aquello que pueden o no publicar y responden los comentarios de los lectores.

2. Construcción del corpus de blogs

Para la construcción del corpus, primero creamos una lista de blogs aptos para ser incluidos en el corpus, y usamos la información publicada en el perfil del autor para extraer información relevante sobre los aprendices como el sexo, la edad, el lugar de

3. Mizumoto *et al*. 2011 crearon un corpus de aprendices de japonés a partir de la red social para el aprendizaje de lenguas extranjeras Lang-8 (en la que los aprendices de varias lenguas escriben y se corrigen entre ellos) con el objetivo de desarrollar un sistema de corrección automática de errores.

residencia, etc. así como sobre su motivación para escribir en español. Después, compilamos los textos y los sometimos a varias etapas de procesamiento hasta ser incluidos en el corpus.

2.1. Búsqueda de blogs y perfil de los autores

Buscamos blogs que cumplieran los requisitos siguientes:

a. Están escritos mayoritariamente en español. Cierto grado de mezcla de lenguas es aceptable, pero excluimos blogs bilingües en los que cada texto está escrito en dos o más lenguas de forma sistemática.

b. La lengua materna de los autores es el japonés. Excluimos blogs escritos por hablantes de español como lengua heredada.

c. Son escritos por una única persona. Excluimos blogs de clase en los que no podemos determinar la identidad de la persona que escribe el texto.

d. No tienen fines comerciales. Excluimos blogs escritos para captar clientes en un negocio en línea.

e. Contienen texto genuino. Excluimos blogs que copian noticias de periódicos en español u otras fuentes de forma sistemática.

Aplicando estos criterios obtuvimos una lista de 48 blogs escritos por 43 autores, alojados principalmente en los dominios de Blogger y Wordpress. Gracias a la información publicada en el perfil de cada blog, asignamos alguna información relevante a cada autor, como el sexo, la localización y otro tipo de información personal. Dos terceras partes de los autores (29/43) son mujeres y una tercera parte son hombres (14/43). Casi la mitad están viviendo en países hispanos en la última actualización (16 en España, 3 en México y 1 en Costa Rica), para estudiar español como estudiantes universitarios (por uno o dos semestres) o son residentes por otros motivos, generalmente por un periodo de tiempo más largo (entre 2 y 18 años). Entre los que viven en Japón (23/43), muchos han visitado o vivido en países hispanos. El blog más antiguo empieza en 2004 y el más reciente en 2015. Los blogs más activos comenzaron entre 2010 y 2012.

Lamentablemente, la información que podemos obtener del perfil de los autores del blog o a partir de la lectura de las primeras entradas es limitada. En este sentido, a diferencia de los corpus de aprendices tradicionales, disponemos de poca información sobre la biografía lingüística del aprendiz: simplemente sabemos que son hablantes de japonés como primera lengua, tienen conocimientos variables de inglés (el estudio de esta lengua es obligatorio en varias etapas del sistema educativo) y su nivel de español va desde principiante a avanzado. El hecho de no tener información fiable sobre sus conocimientos de otras lenguas ni sobre su aprendizaje del español (cuánto tiempo lo han estudiado, dónde, etc.) es una de las principales limitaciones de este tipo de textos.

2.2. Recopilación de textos y post-procesamiento

A partir de la lista de los 48 blogs, extrajimos una lista de URLs, una para entrada (texto) del blog, en total 2.701. A principios de 2015 descargamos los archivos HTML correspondientes, que fueron sometidos a un proceso de post-procesamiento para que estos puedan ser incluidos en el corpus. Las etapas de post-procesamiento son las siguientes:[4]

1. Conversación de los archivos a la codificación UTF-8.

2. Selección del cuerpo del texto y exclusión de otras partes de la entrada como el encabezamiento y los comentarios (que son escritos por los lectores).

3. Eliminación del etiquetaje HTML.

4. Exclusión de contenido irrelevante (*boilerplate*) como figuras, contenido multimedia, elementos de navegación, recomendaciones, repetición de títulos, etc.

5. Selección de una extensión mínima para los textos: los textos con menos de 30 palabras fueron excluidas.

6. Identificación de lengua. A pesar de que seleccionamos solo blogs escritos mayoritariamente en español, algunos autores escriben ocasionalmente algunas entradas en otra lengua (sobre todo japonés o inglés), o la misma entrada en dos o más lenguas. Excluimos entradas con menos del 60% de contenido en español con la herramienta ngramj[5].

3. Resultados

A partir de la lista inicial de 2.701 textos, obtenemos 2.125 textos aptos para ser incluidos en el corpus, como muestra la tabla 1.

Tipo de post-procesamiento	Nº de textos resultante
Archivos en HTML	2.701
Conversión a UTF-8	2.700
Selección del cuerpo del texto	2.685
Eliminación de etiquetas HTML	2.682
Eliminación de contenido irrelevante	2.177
Selección de una extensión mínima	2.177
Identificación de lengua	2.125

Tabla 1. Número de textos resultantes tras cada post-procesamiento

4. Para más detalles sobre la construcción de *web corpus*, ver Schäfer y Bildhauer 2013.
5. http://ngramj.sourceforge.net/

Finalmente, el corpus de blogs contiene 2.125 textos, unos 737.396 *tokens* o 625.343 palabras. Ha sido etiquetado automáticamente con información morfosintáctica con Freeling v4 (Padró y Stanislovsky 2012), para poder estudiar sus propiedades lingüísticas con más detalle, y es consultable en la plataforma Sketch Engine (Kilgarriff *et al.* 2014), como muestra la figura 1. Es esperable que la etiquetación automática contenga frecuentes errores al ser aplicada sobre textos de aprendices, pero consideramos que ese margen de error es aceptable para un estudio preliminar (Valverde 2011).

Figura 1. Interfaz de consulta del corpus en la plataforma Sketch Engine. Consulta de ejemplos del verbo *gustar* en cualquier de sus formas (1.125 ocurrencias)

3.1. Motivaciones

Las motivaciones para escribir un blog son variadas y en la literatura se han propuesto varias clasificaciones (Nardi *et al.* 2004). Curiosamente, la motivación principal de nuestros aprendices no es mejorar su nivel de español, como cabría esperar por su condición de hablantes no nativos. Solo tres autores explícitamente destacan que esta es su motivación principal, aunque algunos más puedan tener este objetivo sin decirlo abiertamente. Teniendo en cuenta solamente las motivaciones expresadas en su perfil personal, los aprendices escriben para documentar su vida (20/43) (1), explicar temas de Japón a los lectores de otros países (17/43) (2), mejorar su español (3), discutir sobre temas de actualidad (4) y socializar (5).

(1) Gracias por visitar aqui!soy japonesa y vivo en Osaka.Escribo mi vida para no olvidarla y practicar espanol.

(2) A partir de hoy, voy a empezar este blog para que muchos hispanohablantes conozcan y tengan interés a mi país.

(3) Soy japonés y estoy pensando emigrar a Sudamérica. Hago este blog para que me aprenda e. Por eso hay muchas faltas de las gramáticas, las expresiones, las ortografías, etc.español.

(4) Soy Natural de Iwaki (Fukushima, la provincia más afectada por el terremoto) y voy a publicar mis pensamientos, noticias y opiniones de personas que han sufrido y sufren esta pesadilla en mi pueblo.

(5) Tengo 17 años, soy Japonesa pero vivo en España. Quiero hacer amigos.

3.2. Temas

Los blogs que forman nuestro corpus tratan principalmente sobre uno de los siguientes temas: la vida personal de los autores (23/48), la cultura japonesa (12/48), la gastronomía (8/48), las lenguas (3/48) y el accidente nuclear de Fukushima (2/48).

En concordancia con las motivaciones principales de los aprendices, estos escriben principalmente sobre su vida diaria, en Japón u otro país hispano, y sobre temas de cultura japonesa.

3.3. Propiedades lingüísticas de los textos

Los textos procedentes de blogs se caracterizan por ser un género a medio camino entre la oralidad y la escritura. Además, los blogs escritos por aprendices, a diferencia de lo que ocurre en los textos escritos en el centro de enseñanza, presentan un uso frecuente de palabras en la lengua materna, el japonés, con distintas funciones, como veremos a continuación.

3.3.1. Oralidad y escritura

Los blogs, ya sean escritos por hablantes nativos o por aprendices, presentan rasgos tanto del registro escrito como del registro oral (Nilsson 2003 y Herring *et al.* 2005).

Por una parte, como en el género escrito, los blogs están condicionados por el espacio que ocupan y son asíncronos, lo que facilita que puedan contener construcciones características de la escritura, como vocabulario especializado (6) y oraciones complejas (7), y puedan ser revisados, tanto antes como después de su publicación.

(6) Se discutirán los poderosos del tema de todo el planeta sobre la *descontaminación* y el seguimiento de la salud de la población.

(7) Hoy os hablaré de“Kagami biraki” (鏡開き) es una ceremonia tradicional japonesa *cuyo* nombre significa literalmente “abrir el espejo”.

Por otra parte, como en el género oral, los blogs son dinámicos (la primera página, que contiene la entrada más reciente, cambia a menudo, mientras que las entradas

antiguas son archivadas), promueven la inmediatez (los textos pueden ser publicados poco después del evento que narran) e intentan reflejar algunos de los matices del habla mediante emoticonos o interjecciones.

En nuestro corpus, es curioso el uso de emoticonos, tanto de estilo occidental (8) como de estilo oriental (9) y 2channel (10), estos dos últimos con caracteres del alfabeto japonés;[6] así como el empleo de interjecciones (11), argot (*tío, chaval, guay, chulo, molar, cojones, cabrón*, etc.) (12, 13) y frases inacabadas (14).

(8) Mi hermano mide 198cm, y lo come desde por la mañana XD

(9) Tanto tiempo no he escrito!! Perdon por vaga!!! m(_ _)m

(10) Hoy era 2años de nuestro matorimonio.(^^)Y☆Y(^^)

(11) Hoy, hace buen tiempo. Esta bien. Esta mañana he hecho la colada cuarto veces. ¡Uf! Estoy un poco cansado.

(12) Un *tío* que había viajado fuera de Japón y le habían gustado los baños arabes montó una sauna en Ginza donde te podías meter en una sauna privada y unas chicas te daban masaje.

(13) Hoy en día , está de la moda el kimono entre los jovenes de Japón, hay muchos accesarios del kimono, y la manera de vestir el kimono también es más *chula* y *guay* como con calcetines (TABI), las botas , la estola y la mochila ...

(14) Ya no está abierto la tienda ni super.....Las comidas que tenia en la casa son sólo cerdo, pasa, nueces, naranja..... Pues por eso he añadido todo con un poco de aceite de oliva y la mantequilla.

Por último, como se ha observado en muchas lenguas, la proporción de sustantivos, adjetivos y preposiciones es mayor en los textos formales, mientras que en los textos informales es mayor la proporción de pronombres personales, adverbios y verbos léxicos (Heylighen y Dewaele 1999). En nuestro corpus, como era esperable, la distribución de frecuencias de estas clases de palabras es la propia de los textos informales. Como muestra la tabla 2, los verbos léxicos constituyen el 14,8% de palabras, los adverbios el 5% y los pronombres personales el 2,7%, cifras invariablemente superiores en comparación con otros tres corpus: un corpus de textos académicos escritos por estudiantes universitarios japoneses en su último año de la carrera de español (APU), de unas 60.000 palabras (Valverde 2015), un corpus de textos académicos escritos por estudiantes universitarios –hablantes nativos– en México (UNAM),[7] de unas 250.000 palabras, y el corpus esTenTen (Jakubíček *et al*. 2013), un web corpus de casi diez mil millones de palabras.

6. El emoticono del ejemplo 9 representa la inclinación de una persona como signo de respeto o disculpa (la letra eme representa las manos y el espacio entre paréntesis, la cabeza inclinada).
El emoticono del ejemplo 10 representa a dos personas brindando con una copa (la letra i griega) en la mano.
7. Este corpus es parte del corpus CLAE (2009), http://www.lenguajeacademico.info/

Tipo de corpus	Corpus de aprendices				Corpus nativo			
Tipo de texto	**Blog**		**Académico**		**Académico**		**Web**	
Nombre	Blog corpus		APU		UNAM		esTenTen	
Frecuencia	f_i	%	f_i	%	f_i	%	f_i	%
Verbos léxicos	92.306	14,8	7.307	11,6	29.648	11,5	1.155.935.910	12,2
Adverbios	31.341	5,0	1.968	3,1	7.917	3,1	285.060.011	3,0
Pron. personal	16.648	2,7	630	1,0	2.750	1,1	120.290.402	1,3
Total palabras en el corpus	625.343	100	63.069	100	258.597	100	9.497.402.122	100

Tabla 2. Frecuencia absoluta y relativa de verbos léxicos, adverbios y pronombres personales en dos corpus de aprendices (corpus de blogs y APU) y dos corpus nativos (UNAM y esTenTen)

3.3.2. Uso de palabras extranjeras

El cambio de código, esto es, la alternancia entre dos o más lenguas en el discurso de los hablantes bilingües, es común en los blogs escritos por dichos hablantes (Montes-Alcalá 2007). En nuestro corpus de blogs apenas encontramos cambio de código propiamente dicho –de hecho, excluimos de nuestro corpus los blogs escritos por hablantes de español como lengua heredada, que probablemente poseen un mayor dominio del español, y son por lo tanto más propensos a usar cambio de código–. En cambio, encontramos frecuentemente palabras o expresiones escritas en japonés, en ocasiones al lado del equivalente español. Esta mezcla de lenguas se da con varias funciones, como veremos más adelante.

La mezcla de ambas lenguas, español y japonés acarrea también la mezcla de cuatro sistemas de escritura distintos, lo que complica el tratamiento automático de los textos:

a. El alfabeto latino, usado para escribir en español o para transcribir el japonés, a menudo en mayúsculas, como en el ejemplo 15, *dashi, konbu* y *katsuobushi*.

b. Los kanjis japoneses (caracteres logográficos de procedencia china). En el ejemplo 15, 昆布 y鰹節, que se leen ***konbu*** y ***katsuobushi*** respectivamente.

c. El silabario hiragana y el silabario katakana. En el ejemplo 15 aparece だし en el silabario hiragana, que se lee *dashi*.

(15) Lo más importante para la comida japonesa es “Dashi だし”.
- ¿Qúe es Dashi?
Es un caldo japonés. Usamos para casi todos los platos japoneses.
=== Ingredientes ===
- Alga (se llama KONBU 昆布) 20 gr. (unos 10 cm. más o menos)

- Bonito seco (se llama KATSUOBUSHI 鰹節) 30 gr.
- Agua 100 ml. y 1000 ml.

A pesar de que el texto en español es escrito mayoritariamente con el alfabeto latino, es frecuente encontrar no solo palabras en japonés, sino también signos de puntuación,[8] caracteres especiales (como ☆) e incluso mezcla de caracteres de ancho medio y ancho completo (16) en la misma frase.

(16) "Ｅｌ　ｔｅａｔｒｏ　en la parte de atrás de mi es el sitio que bailar　el Nou."

Es interesante que los aprendices incluyan a menudo palabras en kanji y en kana, al lado de su transcripción al alfabeto latino, teniendo en cuenta que los lectores probablemente no pueden leer esos símbolos. La inclusión de todas las formas escritas de una palabra puede ser explicada probablemente por la naturaleza dual de los blogs: al ser publicaciones a la vez privadas y públicas, el texto debe ser cómodo de leer tanto para el que lo escribe (que prefiere usar el sistema de escritura japonés para escribir las palabras japonesas) como para el que lo lee (que necesita la traducción y la transcripción).

Los aprendices usan el japonés dentro de los textos en español con varias funciones. En primer lugar, lo más frecuente es que usen palabras para expresar conceptos ligados a su cultura, para los cuales no existe una traducción equivalente al español, por ejemplo los ingredientes de una receta (15), inventos o productos japoneses (17), festividades o tradiciones (18) y nombres propios (19). Por lo tanto, el uso ocasional de palabras japonesas en los textos escritos en español no debe ser interpretado como el resultado de una falta de dominio lingüístico, sino como el resultado de la ausencia de un equivalente en la lengua meta. Nuestros aprendices se encuentran entre dos lenguas y dos culturas, y necesitan usar ambas para poder expresarse adecuadamente.

(17) Por último, en Japón existe un sentimiento de vergüenza a que los demás escuchen los sonidos que se producen mientras está utilizando el water. Así existe un sistema llamado OTOJIME(REINA DEL SONIDO). Si pulsas este botón, mientras utilizas el water, se produce un sonido de corriente de agua durante unos 15 segundos.

(18) Dentro de un mes, el 5 de mayo, hay una fiesta para los niños en Japón.Se llama "Día de los niños (kodomo no hi こどもの日)".

(19) Pero en esta semana he encontrado una poesía. Se llama 「朝のリレー」(Relevos de la mañana), por 谷川俊太郎（Tanigawa Shuntaro）Cuando leía una revista japonesa que me mandó una amiga mía, encontré un artículo del poeta y poesía.

En segundo lugar, usan palabras japonesas en contextos metalingüísticos, cuando hablan sobre la lengua japonesa, especialmente sobre el significado de algunas palabras (20) y proverbios (21).

8. Aunque a primer vista puedan parecer los mismos, en japonés se usan signos de puntuación ligeramente distintos del alfabeto latino, como comillas (“”), puntos (。), puntos suspensivos (...) , paréntesis, etc.

(20) ナウい(Naui) Adjetivo que significa “a la moderna”. Procede de la palabra inglesa “now” y “i” que es el final de una palabra para hacer un adjetivo japonés. ej) “Su vestido es muy Naui!!” El japonés que usa esto es la persona de la mayor edad.

(21) En japonés hay una frase hecha “Ocha wo Nigosu (literalmente, “enturbiar el té”)”, que significa “engañar”.

En tercer lugar, aparecen fragmentos en japonés en las citas, dentro del discurso directo (22) o indirecto (23).

(22) Cuando terminamos, me levanté y dije “*muchas gracias por la entrevista* 本日はお時間をお取りいただきありがとうございました”, haciendo una reverencia. Y otra vez delante de la puerta hice otra reverencia.

(23) Según la publicación por la Cruz Roja de Japón, al 16 de mayo, se juntaron las donaciones de 1,925億6,316万9,033円 (192.563.169.033 yenes) que equivalen a 1.673 millones de euros para asistir a los damnificados por el gran terremoto y el tsunami del 11 de marzo.

En cuarto lugar, la repetición de la misma frase en las dos lenguas, español y japonés, sirve para dar énfasis, a menudo al principio de la entrada, en el título (24), o al final, en la despedida (25).

(24) Mi cumpleaños　私の誕生日
(25) Qué durmais como marmotas. ぐっすりおやすみください

Por último, algunas expresiones en japonés se repiten con frecuencia al final de las oraciones. Los aprendices no proporcionan la traducción de estas palabras y podrían considerarse un ejemplo de cambio de código. Se trata principalmente de interjecciones u otro tipo de expresiones breves que pertenecen a la función expresiva del lenguaje, ya que su objetivo principal es revelar el sentimiento del emisor (26-28).[9]

(26) Paulaaa, *gomennn* tambien estan tus fotos de la fiesta de tu cumpleanos :D *omatase*...

(27) ¡Es muy curioso y me encanta! *KAWAII*--------!!!!!

(28) Una amiga de Blugaria hizo fotografias.Bien, *ne*^^

4. Conclusiones

El uso de la web como corpus en la creación de corpus de aprendices nos da acceso inmediato a una gran cantidad de textos en formato electrónico, que han sido produ-

9. En el ejemplo 26, *gomenn* se usa para pedir disculpas y *omatase* se usa para disculparse por la espera.
En el ejemplo 27, *kawaii!* equivale a ¡qué bonito! en español.
En el ejemplo 28, *ne* es una partícula que se usa para confirmar lo dicho con anterioridad, equivalente a ¿verdad? o ¿sí? en español.

cidos en unas condiciones distintas de las habituales en los centros de enseñanza: los aprendices escriben sus entradas del blog sin limitaciones de tiempo, extensión, estilo o temas.

Los textos publicados en los blogs presentan una mezcla de rasgos propios de la lengua escrita y la lengua oral y nos sirven para conocer mejor las motivaciones y temas de escritura preferidos de sus autores. Las principales motivaciones de los aprendices japoneses de español para escribir un blog son dejar constancia de sus vivencias, de la misma forma que en un diario personal, y explicar aspectos de la cultura japonesa a los lectores extranjeros. Así, los temas preferidos son la vida diaria y la cultura japonesa.

Otra de las ventajas de este tipo de textos es que han sido escritos en soporte electrónico por los propios aprendices, a diferencia de los textos escritos en el centro de enseñanza, generalmente manuscritos, entre otros motivos por la falta de dominio del sistema de escritura a ordenador en español –sobre todo en los niveles iniciales-, ya que recordamos que el japonés usa un sistema de escritura distinto compuesto por ideogramas de origen chino y dos silabarios. Además, a diferencia de los blogs escritos por hablantes nativos de español y de las tareas de clase que solemos leer los profesores, en los blogs de aprendices destaca el uso frecuente de palabras en la lengua materna, el japonés, principalmente para referirse a conceptos de su cultura para los que a veces no existe una traducción precisa al español.

En cuanto a las desventajas de este tipo de textos, está el hecho de que disponemos de un escaso número de aprendices (menos de cincuenta) y tenemos poco control sobre el perfil de estos. Además, el número de textos por autor es desigual, y suele estar relacionado con su nivel de español: los aprendices con mayor dominio del español escriben más textos y textos más largos que los aprendices con menor dominio. Por último, sería deseable poder dar libre acceso al corpus a la comunidad investigadora; sin embargo, los derechos de autor permiten la consulta pero no la distribución (a menos que consigamos el permiso expreso de los autores).

Por todo ello, nuestra investigación futura consistirá en la creación de un corpus de aprendices que reúna textos de distinto tipo y producidos en distintos contextos (tareas de clase, exámenes y también publicaciones electrónicas como los blogs), de forma que el usuario del corpus pueda seleccionar el tipo de texto más adecuado para su investigación, ya sea en el campo de la adquisición de segunda lenguas, la enseñanza de lenguas extranjeras o el procesamiento del lenguaje natural. Mientras que en el campo de la adquisición de segundas lenguas es esencial disponer de información sobre el perfil de los aprendices, esa información puede no ser necesaria por ejemplo en ciertas áreas del campo del procesamiento del lenguaje natural.

Agradecimientos

Esta investigación ha sido parcialmente financiada por *Grant-in-Aid for Scientific Research (Start-up)* de la Japan Society for the Promotion of Science (17H07270).

Referencias Bibliográficas

ALONSO-RAMOS, M. (2016): *Spanish Learner Corpus Research: Current trends and future perspectives*. Studies in Corpus Linguistics, 78, John Benjamins.

EFIMOVA, L., y DE MOOR, A. (2005): "Beyond personal webpublishing: An exploratory study of conversational blogging practices", *Proceedings of the 37th Annual HICSS Conference*. Big Island, Hawaii.

GILL, A. J., NOWSON, S., y OBERLANDER, J. (2009): *What are They Blogging About? Personality, Topic and Motivation in Blogs*, Association for the Advancement of Artificial Intelligence.

GRANGER, S., GILQUIN, G. y MEUNIER, F. (2015): *The Cambridge Handbook of Learner Corpus Research*, UK, CUP.

HERRING, S., SCHEIDT, L., SABRINA BONUS, S., y WRIGHT, E. (2005): "Weblogs as a bridging genre", *Information, Technology & People*, 18 (2), 142-171.

HEYLIGHEN, F., DEWAELE, J. M. (1999): *Formality of language: definition, measurement and behavioral determinants*. Technical report, Free University of Brussels.

JAKUBÍČEK, M., KILGARRIFF, A., KOVÁŘ, V., RYCHLÝ, P., y SUCHOMEL, V. (2013): "The TenTen Corpus family", *7th International Corpus Linguistics Conference*, Lancaster.

KILGARRIFF, A., BAISA V., BUŠTA, J., JAKUBÍČEK, M., KOVÁŘ, V., MICHELFEIT J., RYCHLÝ, P., SUCHOMEL, V. (2014): "The Sketch Engine: ten years on", *Lexicography*, 1–30.

LEACOCK, C., M. CHODOROW, M. GAMON, J. TETREAULT (2014): *Automated Grammatical Error Detection for Language Learners*, Morgan & Claypool Synthesis Lectures Human Language Technologies, Toronto.

LOZANO, C. (2009): "CEDEL2: Corpus Escrito del Español como L2", *Applied Linguistics Now: Understanding Language and Mind*, Almería, Universidad de Almería, 197–212.

LU, H. C. (2010): "An annotated Taiwanese learners' corpus of Spanish, CATE", *Corpus linguistics and Linguistic Theory* 6 (2), 297-300.

MANCERA, A. M. C., MARTÍNEZ, I. P., CANALES, A. B., FERNÁNDEZ, L. C., GRANDA, J. F. S. (2001): "Corpus para el análisis de errores de aprendices de E/LE (CORANE)", *Actas del XII Congreso Internacional de ASELE: tecnologías de la información y de las comunicaciones en la enseñanza de la E/LE*, 527–534.

MENDIKOETXEA, A. (2014): "Corpus-based research in second language Spanish", *The Handbook of Spanish Second Language Acquisition*, Blackwell, UK.

MONTES-ALCALÁ, C. (2007): "Blogging in two languages: code-switching in bilingual blogs", *Selected Proceedings of the Third Workshop on Spanish Sociolinguistics*, Somerville, MA: Cascadilla Proceedings Project, 162-170.

NARDI, B. A., SCHIANO, D. J., GUMBRECHT, M., SWARTZ, L. (2004): "Why We Blog", *Communications of the Association for Computing Machinery*, 41-46.

NILSSON, S. (2003): *The function of language to facilitate and maintain social networks in research weblogs*. Dissertation, Umea Universitet, Engelska lingvistik.

PADRÓ, L. y STANILOVSKY, E. (2012): "FreeLing 3.0: Towards Wider Multilinguality", *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)* ELRA. Istanbul, Turkey.

ROJO, G., y PALACIOS, I. (2016): "Learner Spanish on computer", *Spanish Learner Corpus Research: Current trends and future perspectives,* John Benjamins.

SCHÄFER, R. y BILDHAUER, F. (2013): *Web Corpus Construction*, Synthesis Lectures on Human Language Technologies, Morgan&Claypool Publishers.

VALVERDE, M. P. (2011): "An evaluation of part of speech tagging on written second language Spanish", *Lecture Notes in Computer Science*, vol. 6609, Springer Berlin Heidelberg, 214-226.
VALVERDE, M. P. (2015): "Frecuencia de uso de palabras gramaticales en textos académicos: Comparación de un corpus aprendices de ELE con tres corpus de referencia", *Hispánica*, 59, 127-154.