DISEÑO Y CREACIÓN DE UN CORPUS DE APRENDICES DE ESPAÑOL EN JAPÓN (CELEN)

Pilar Valverde
Universidad Kansai Gaidai, Japón

RESUMEN

Presentamos CELEN, el primer corpus de aprendices dedicado en exclusiva al análisis de la expresión escrita de los estudiantes japoneses de español. Se trata de un corpus con vocación pedagógica, concebido como una suma de varios subcorpus locales, que son una muestra representativa de los textos que los estudiantes escriben durante un año académico en las clases de ELE de una universidad determinada. En este artículo se describe el primer subcorpus, correspondiente a la Universidad Kansai Gaidai, formado por 1.840 textos de 459 aprendices, unas 140.000 palabras. Se trata de un corpus abierto que está siendo ampliado con textos de otras instituciones. Está disponible para descarga bajo una licencia CC-NY-NC y puede ser consultado en línea en la plataforma Sketch Engine.

Palabras clave: corpus de aprendices, español en Japón, aprendices japoneses de español.

ABSTRACT

We present CELEN, the first learner corpus geared to the analysis of the written expression of Japanese students of Spanish. It is a corpus with a pedagogical aim, conceived as a sum of several local subcorpora, which constitute a representative sample of the texts that students write during one academic year in the Spanish classes of a given university. This article describes the Kansai Gaidai University subcorpus, made up of 1,840 texts from 459 learners, about 140,000 words. CELEN is an open corpus and is being extended with texts coming from others institutions. It is available for download under a CC-NY-NC license and can be consulted online in the Sketch Engine platform.

Keywords: learner corpora, Spanish in Japan, Japanese learners of Spanish

1. EL ESPAÑOL COMO LENGUA EXTRANJERA EN JAPÓN

En Japón el español es una lengua de estudio minoritaria, con una popularidad semejante a la de otras lenguas europeas como el francés o el alemán, y muy por detrás del coreano, el chino y la lengua extranjera por excelencia, el inglés. El Instituto Cervantes en su último informe (2018) calcula que hay unos 60.000 estudiantes de español en el país y, en la clasificación de países por número de estudiantes, Japón ocupa el puesto número 20 —en el número 1 se encuentra Estados Unidos (más de ocho millones de estudiantes) y en el 2, Brasil (más de seis millones)—.

En cuanto a los contextos de instrucción, en la enseñanza reglada predomina el nivel universitario, seguido a mucha distancia por la secundaria y bachillerato[1], mientras que la presencia del español en la escuela primaria es prácticamente nula. En la enseñanza no reglada encontramos las academias de idiomas[2], entre las que podemos incluir el Instituto Cervantes de Tokio[3], los centros culturales y los que estudian español de forma independiente (véase Sanz *et al.*, 2015 y Ugarte, 2012).

Dado el reducido número de estudiantes y su concentración en ciertos contextos de enseñanza, por el momento el corpus que presentamos aquí recoge exclusivamente textos del ámbito universitario, donde suelen distinguirse dos modalidades de estudio: el español como asignatura de segunda lengua extranjera y el español como carrera.

En el contexto universitario, los estudiantes de carreras muy diversas tienen que estudiar durante un año académico (30 semanas) una segunda lengua extranjera, esto es, una lengua extranjera distinta del inglés, y el español es una de las opciones ofrecidas[4]. Normalmente se imparten 2 clases (de una

[1]Según los datos del Ministerio de Educación, Cultura, Deporte, Ciencia y Tecnología de Japón (MEXT), de 2003 a 2009, el número de centros de bachillerato que impartían otras lenguas extranjeras además del inglés pasó de 454 a 2.027, y el español se ofrecía solo en un centenar de ellos. Asimismo, el número de estudiantes de bachillerato que estudiaban español se triplicó de 2003 a 2009, siendo 2.763 en 2009 (Ugarte, 2012).

[2]Según los datos del Ministerio de Economía, Comercio e Industria de Japón, en 2005 había 1.144 academias y el 12,5% (esto es, unas 150) ofrecía clases de español (Ugarte, 2012).

[3]En 2016 tenía 914 alumnos en sus cursos (Instituto Cervantes, 2018), una cifra semejante a la de los centros de Austria, Bélgica, Bulgaria, Túnez o República Checa.

[4]En 2009 el español ocupaba el sexto lugar en la lista de idiomas que ofrecen las universidades, tras el inglés, el chino, el francés, el alemán y el coreano.

hora y media de duración cada una) por semana, una a cargo de un profesor japonés y otra a cargo de un profesor hispanohablante. Se trata por lo tanto de un curso de corta duración (90 horas) en el que generalmente se prioriza el conocimiento teórico, razón por la cual al terminar el curso, la mayoría de estudiantes tiene un nivel básico y no superaría un examen de nivel A1.

En cuanto al contexto universitario especializado, el español puede estudiarse de forma intensiva, como parte fundamental de la licenciatura, en un departamento de español. Las clases de esta lengua (5 o más a la semana, esto es, 7,5 horas) suelen ser obligatorias durante los dos primeros años y optativas en los dos últimos, durante los cuales se priorizan los estudios de área (economía, política, relaciones internacionales, historia, literatura, etc.). Como resultado, al terminar la carrera los estudiantes alcanzan un nivel que puede ir desde A2, en el caso de los estudiantes de mediano rendimiento que se quedan en Japón, hasta B2, en el caso de los alumnos más aventajados que realizan estancias en el extranjero.

De las 780 universidades del país, más de 200 ofrecen el español como lengua extranjera optativa (Ugarte, 2012), mientras que solo una quincena ofrece la especialidad de español[5]. Esto significa que la gran mayoría de estudiantes universitarios de español se encuentran por debajo del nivel A1 y solo unos 4500 estudian "Filología Hispánica" (Sanz, 2013).

En el corpus que presentamos aquí se incluirán textos procedentes de ambos contextos, pero prestando especial atención al contexto especializado, con el objetivo de recoger textos correspondientes a varios niveles de dominio, al menos desde A1 hasta B2 del MCER.

En cuanto al perfil de los estudiantes, la gran mayoría no tiene conocimientos previos de español al empezar sus estudios universitarios, ni contacto con esta lengua fuera de clase, y apenas usará el español en su trabajo después de graduarse, pues en este país la especialidad cursada es prácticamente irrelevante para la vida laboral. Según los resultados de la encuesta realizada por el grupo GIDE (2012) en 40 universidades, la mayoría estudia español porque es una de las lenguas más habladas del mundo y porque le interesa la cultura del mundo hispano (la música, el deporte, el baile, etc.), y su principal objetivo es usar la lengua para comunicarse con la gente cuando viajen a países de habla española.

[5]En algunas universidades también existe la posibilidad de cursar cuatro o más clases de español semanales, pero no con la categoría de especialidad.

Como la mayoría de universidades del país —tres de cada cuatro— son privadas[6], los métodos de enseñanza, la formación del profesorado y las directrices de cada una varían considerablemente. Mientras que para graduarse en algunas se requiere alcanzar un nivel suficiente para escribir un resumen de la tesina de graduación y superar una entrevista sobre la misma en español, en otras se exige solamente que los estudiantes hayan superado las asignaturas obligatorias y obtenido los créditos necesarios, independientemente de su nivel de dominio del español en el momento de graduarse. Para saber más sobre la situación de la enseñanza de ELE en el contexto universitario japonés y las propuestas de reforma, véase Sanz (2013), Sanz et al. (2015), Moreno (2015) y Escandón (2017).

2. LOS CORPUS DE APRENDICES DE ESPAÑOL

El desarrollo de corpus de textos escritos por aprendices de una lengua extranjera, denominados *corpus de aprendices*, empezó en los años ochenta del siglo XX. El primer corpus de aprendices a gran escala fue el *International Corpus of Learner English* (ICLE) (Granger *et al.*, 2009), coordinado desde la Universidad de Lovaina, cuya primera versión fue publicada en CD-ROM en 2002 y que en la versión 2, publicada en 2009, contiene 4,5 millones de palabras procedentes de aprendices de inglés con once lenguas maternas distintas.

En el ámbito hispánico el desarrollo de corpus de aprendices se encuentra todavía en una fase inicial (Alonso-Ramos, 2016). Si acudimos al repositorio digital de recursos lingüísticos del proyecto europeo CLARIN[7], en el apartado dedicado a corpus de aprendices[8], encontramos 24 corpus monolingües de aprendices de nueve lenguas (inglés, finés, sueco, alemán, árabe, checo, francés, húngaro y noruego), pero ninguna de ellas es el español. Tenemos que llegar al apartado de "Corpus que se encuentran fuera de la infraestructura de CLARIN" para encontrar solamente un corpus de español disponible para consulta y descarga, el *Spanish Learner Language Oral Corpus* (SPLLOC)[9], un

[6] Según el MEXT en 2018 había 780 universidades: 604 privadas, 86 nacionales y 90 locales. En línea: http://www.mext.go.jp/b_menu/toukei/002/002b/1403130.htm.

[7] *Common Language Resources and Technology Infrastructure* (CLARIN) es un proyecto europeo que tiene como objetivo desarrollar y reunir recursos tecnológicos en relación al uso y aplicación del lenguaje y de la cultura.

[8] https://www.clarin.eu/resource-families/L2-corpora.

[9] http://www.splloc.soton.ac.uk/.

corpus oral destinado a la investigación sobre la adquisición del español como lengua extranjera.

Existen otros listados en línea más amplios, como el *Learner corpora around the world*[10], mantenido por los autores del ICLE de la Universidad de Lovaina. En dicha lista, en el momento de escribir este artículo, figuran 19 corpus de aprendices de español. Sin embargo, la mayoría de los que aparecen en ella no están disponibles para ser descargados o consultados online por otros investigadores, ya sea porque son de acceso restringido, están todavía en una fase de desarrollo o simplemente quedaron inacabados y nunca fueron publicados. Así pues, a la escasez de corpus se suma la dificultad para acceder a la mayoría de ellos, lo que limita su utilidad.

Si centramos nuestra atención exclusivamente en los corpus escritos, como el que aquí presentamos, y sin tener en cuenta corpus de reducido tamaño compilados para llevar a cabo investigaciones particulares[11], existen en este momento dos que pueden ser consultados libremente, el Corpus Escrito del Español L2 (CEDEL2)[12]y el Corpus de Aprendices de Español como Lengua Extranjera (CAES)[13], y cinco mediante registro o compra[14]. Cuatro contienen datos de aprendices con una lengua materna determinada —el inglés (CEDEL2), el taiwanés (CATE), el italiano (CORESPI) y el portugués (COMET)— y el resto, lenguas maternas variadas.

El primer corpus de aprendices para el español fue CORANE (Cestero et al., 2001), que reúne 957 composiciones escritas por estudiantes de ELE de los cursos de español para extranjeros de la Universidad de Alcalá de los niveles A2, B1, B2 y C1, y de varias lenguas maternas. Los datos fueron obtenidos en el año 2000 y publicados en formato CD-ROM en 2009.

---

[10]https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

[11]Como los publicados en el portal de la revista Linred: http://www.linred.es/corpus.html.

[12]http://cedel2.learnercorpora.com/.

[13]https://galvan.usc.es/caes.

[14]Son los siguientes:

- El corpus *Aprender a Escribir en Lovaina* (Aprescrilov).
- Corpus para el Análisis de errores de aprendices de E/LE (CORANE).
- El Corpus Multilíngue para Ensino e Tradução (COMET).
- El Corpus de Aprendices Taiwaneses de Español (CATE).
- El Corpus del Español de los Italianos (CORESPI).

El corpus de mayor tamaño es *Aprescrilov* (Buyse, 2012), otro de los pioneros, que cuenta con aproximadamente un millón de palabras recogidas desde 2004 en la Universidad de Lovaina.

Le sigue en tamaño CEDEL2 (Lozano y Mendikoetxea, 2013), que desde 2006 recoge textos escritos por hablantes de inglés L1 a través de un formulario electrónico, llegando actualmente a las 800.000 palabras. Este es además el único que permite la descarga de los datos.

El corpus CAES (Rojo y Palacios, 2016), promovido por el Instituto Cervantes, contiene muestras de hablantes de seis lenguas maternas y, a pesar de su reducido tamaño —575.000 palabras en total— fue concebido como un corpus abierto y en el futuro aspira a constituirse en un corpus representativo del español, al estilo del corpus ICLE para el inglés.

Entre todos los corpus publicados, solamente CORANE contiene algunos textos escritos por hablantes de japonés, por lo que podemos afirmar que CELEN, el corpus que presentamos aquí, es el primero dedicado en exclusiva a aprendices japoneses de español.

3. DISEÑO DEL CORPUS

El desarrollo de corpus de aprendices, a diferencia del desarrollo de corpus de nativos, se ve fuertemente condicionado por el campo del que proceden los investigadores implicados y el uso que se le pretende dar. Así, podemos diferenciar principalmente tres tipos de corpus según los destinatarios principales, que pueden ser:

1. Profesionales de la enseñanza como profesores, creadores de materiales didácticos, evaluadores, etc. Se trata del tipo más común, al que pertenecen los corpus CORANE, *Aprescrilov* y CAES.
2. Investigadores sobre adquisición de lenguas, como el corpus CEDEL2 o SPLLOC.
3. Investigadores en el campo del procesamiento del lenguaje natural o la lingüística computacional, como los corpus multilingües de Lang-8 (Mizumoto *et al.*, 2011) y WordReference (Berdicevskis, 2018).

CELEN nace con una vocación pedagógica, ya que su objetivo principal es facilitar la aplicación de la lingüística de corpus a los profesores de ELE que trabajan en Japón. No obstante, uno de los principios fundamentales que

seguimos al desarrollar el corpus es el de favorecer su reutilización por parte de investigadores con intereses diversos. Para conseguirlo se han tomado las siguientes medidas:

1. Cada texto incluye una gran cantidad de información sobre el aprendiz y sobre la situación en la que se escribió, para que cada investigador pueda seleccionar los textos que cumplen ciertos requisitos.

2. El corpus puede ser descargado y manipulado con fines de investigación bajo una licencia CC BY-NC 4.0 y consultado en línea en la plataforma Sketch Engine (véase el apartado 5).

En cuanto a su uso en el ámbito de la enseñanza, si bien es cierto que algunos aspectos de la enseñanza de ELE deben ser revisados y actualizados usando metodología de corpus (o metodología científica en general) (Llorián, 2017), también lo es que la investigación en este campo suele estar alejada de la realidad que se vive en el aula de lenguas extranjeras (Tono, 2016): mientras que los investigadores suelen estar más interesados en crear una descripción del aprendiz "arquetípico", los profesores se muestran más interesados por lo que hacen realmente sus estudiantes particulares en las clases (Seidlhofer, 2002, Mukherjee y Rohrbach, 2006).

Para acortar esa distancia entre el ámbito de la investigación y la realidad del aula creemos que es importante disponer de subcorpus *locales*, esto es, representativos de una institución en particular, para que los profesores puedan investigar sobre las producciones de sus alumnos, en textos que son relevantes en su contexto de enseñanza, y llevar a cabo proyectos de investigación-acción.

Empezamos a recoger datos en la Universidad Kansai Gaidai, donde cerca de 300 estudiantes se matriculan cada año en la carrera de español. Se trata de un corpus abierto y tras la experiencia acumulada en la recogida de datos en esta universidad durante el año académico 2018, esperamos ampliarlo con la colaboración de otras universidades japonesas que ofrecen la licenciatura en español.

En cuanto los criterios lingüísticos, el corpus está formado exclusivamente por textos escritos, recogidos a lo largo de un año académico. Se trata, por lo tanto, de un corpus longitudinal, ya que registra la evolución de un grupo de aprendices a lo largo de un periodo de tiempo, a la manera de un Portafolio. Concretamente, en la Universidad Kansai Gaidai recogimos aproximadamente la mitad de los textos escritos por los estudiantes durante un año académico:

entre 1 y 8 textos por aprendiz (entre 3 y 6 textos del 78 % de aprendices).

En cuanto a las tareas escritas, estas son las que los estudiantes llevan a cabo en el marco de las asignaturas impartidas por profesores nativos como parte de exámenes parciales o finales, actividades de clase o tareas para escribir en casa. Por lo tanto, están basadas en un tema propuesto por el profesor, y la mayoría son guiadas en el sentido de que incluyen algunos puntos que los estudiantes deben incluir en su texto, al estilo de las tareas de expresión escrita de los exámenes DELE, o semi-guiadas, a partir de un tema propuesto por el profesor. Algunas han sido escritas con limitaciones de tiempo, mientras que otras han sido escritas en casa sin restricciones de tiempo. Asimismo, se indica si el aprendiz ha tenido la posibilidad de acceder a ayuda externa (diccionario, Internet, etc.) o no.

Sobre los aprendices, se trata como hemos comentado antes de jóvenes universitarios, ya que la enseñanza de esta lengua es prácticamente inexistente en etapas educativas anteriores y nuestro corpus tampoco contiene por el momento textos procedentes de academias de idiomas, centros culturales u otros contextos de enseñanza. Se incluyen solamente textos de aprendices cuya lengua materna es, al menos, el japonés. Se indica también el nivel de español e inglés según certificados oficiales, si tiene alguno.

El nivel de español de los textos según el MCER es determinado según el nivel de la asignatura en la cual se escribió el texto, aunque este nivel no se corresponda siempre con el nivel "real" de los textos en términos de competencia lingüística, o según el número de meses u horas de clase recibidas. De esta forma, durante el primer año se usa un libro de texto de nivel A1, y durante el segundo año, un libro de nivel A2, aunque el grado de dominio lingüístico de cada estudiante puede variar dependiendo de factores individuales como la motivación o las horas de estudio, entre otros. Los textos del corpus, por lo tanto, reflejan el nivel de dominio que los profesores de ELE encuentran en sus clases.

4. RECOGIDA DE DATOS

4.1 La institución: solicitud al comité de ética

Cada vez más universidades han actualizado sus políticas con respecto a la investigación que involucra el uso de sujetos humanos, y por ello fue necesario entregar una solicitud al Comité Universitario de Evaluación de la Investigación antes de comenzar la investigación, para garantizar la anonimidad de los participantes.

4.2 Los aprendices: cuestionarios y formularios de consentimiento

Asimismo, fue necesaria la colaboración y coordinación de los profesores del departamento de español, ya que se trata de una de las universidades con mayor número de estudiantes de especialidad, casi 300 por curso. Durante las primeras semanas del curso los profesores de las clases de Gramática entregaron un cuestionario y un formulario de consentimiento a los estudiantes de sus clases. En total, 13 profesores recogieron 508 formularios en 27 grupos de estudiantes.

Para el diseño del cuestionario tuvimos en cuenta los datos incluidos en otros corpus de aprendices de español y en corpus de referencia de otras lenguas, como el ICLE (Granger *et al*., 2009) o MERLIN (Boyd *et al*., 2014), para que el corpus pueda adaptarse a las necesidades de la mayor cantidad de investigadores. Mediante el cuestionario obtuvimos información sobre: edad, sexo, año académico, lengua(s) materna(s) del aprendiz, de la madre y del padre, lengua(s) hablada(s) en casa, contactos personales en países de habla española, estancias en países de habla española, edad de inicio en el estudio del español, títulos oficiales de español y de inglés y conocimientos de otras lenguas.

Aproximadamente el 85 % de los estudiantes completaron el cuestionario. El resto no lo hizo porque estuvieron ausentes ese día o prefirieron no participar. Los perfiles de los aprendices que dieron su consentimiento y que completaron los cuestionarios en su totalidad fueron introducidos en la base de datos. Después de descartar aquellos participantes cuya lengua materna no incluye el japonés o de los cuales no recibimos ningún texto durante el curso

académico, el número total de aprendices incluidos en el corpus es de 459.

Aproximadamente dos tercios son mujeres, como suele ocurrir en las facultades de lenguas extranjeras, y tienen entre 18 y 22 años. El 99 % son hablantes monolingües de japonés y el 1 % hablan japonés y otra lengua. Para el 98% de ellos la lengua materna de la madre, la del padre y la lengua hablada en casa es el japonés, y solo tres viven en una familia con algún miembro cuya lengua materna es el español. Aproximadamente el 80 % pertenece al programa regular, y el resto al programa intensivo (en el que a partir del segundo año reciben más clases de lengua). El 45 % son estudiantes de primer año, el 45 % de segundo, el 9 % de tercero y el 1 % de cuarto. El 90 % empezó a estudiar español a partir de los 18 años. Tres cuartas partes no tienen contactos personales (familia, amigos, etc.) en países de habla hispana y el 88 % no ha estado nunca en un país de habla hispana. Entre el 12 % que sí ha estado, la mayoría ha estado menos de un mes, y los países preferidos son España, Perú y México. En cuanto al nivel de español certificado, un 13 % tiene algún título de español: el más popular es el *Supeingo Kentei (Evaluación del conocimiento de la lengua española)*, organizado por la Sociedad Hispánica del Japón y que tiene seis niveles, desde el 6, el más básico, hasta el 1, el más avanzado. Los niveles más comunes entre los aprendices son el 4 (35 %), 5 (27 %) y 6 (28 %). En lo que respecta al nivel de inglés, el 68 % manifiesta tener uno o más títulos (EIKEN, TOEIC, TOEFL ITP, GTEC u otros), y en su equivalencia con el nivel del MCER lo más común es el nivel B1 (el 56 %), seguido por A2 (30 %) y A1 (9 %).

4.3 Los textos

Nueve profesores a cargo de las clases de Conversación de primer y segundo año del programa regular recopilaron dos textos por semestre (de los cuatro que escriben en total como parte de los exámenes parciales), en total cuatro durante el año académico, en 21 grupos de estudiantes. Cinco profesores de las clases del programa intensivo de segundo a cuarto año recopilaron varios textos en seis grupos, dependiendo de la disponibilidad del profesor. El 77 % de textos proceden de exámenes parciales, el 17 % de tareas, el 4 % actividades de clase y el 2 % de exámenes finales, por lo que el 78 % fue escrito sin acceso a ayuda externa como el libro de texto, el diccionario o Internet.

En total el corpus contiene 1.840 textos, unas 140.000 palabras, y cada texto está relacionado con un gran número de metadatos sobre las características del

aprendiz y del texto en cuestión: fecha, nivel de español de la clase (Apéndice, Figura 1), curso, semestre, meses de clase cursados hasta el momento de escribir el texto, macrofunción predominante (Apéndice, Figura 2), género textual (Apéndice, Figura 3), título, destinatario (Apéndice, Figura 4), autenticidad, noción específica predominante[15] (Apéndice, Figura 5), libro de texto y unidad que se ha estudiado antes de escribir el texto, modo de asignación del tema, situación en la que se escribió el texto, límite de tiempo, acceso a ayuda externa, número de palabras requerido y medio de escritura.

La mayoría de textos fueron escritos a mano por los aprendices, y por lo tanto han sido transcritos con ayuda de OCR o manualmente, revisados, anonimizados e ingresados en la base de datos. Durante la transcripción se ha intentado reflejar lo más fielmente posible el original, incluyendo los errores. Cuando hay ambigüedad en el texto, como suele hacerse en otros corpus de aprendices, evitamos introducir errores de más y hacemos una asunción positiva, esto es, suponemos que el aprendiz no cometió un error. Solo hemos modificado el original para ocultar la información personal que pueda aparecer: nombres propios, números de teléfono, direcciones postales, direcciones de correo electrónico, etc.

5. CONCLUSIONES Y TRABAJO FUTURO

El corpus puede ser consultado en Sketch Engine[16], una herramienta para consultar y crear corpus usada por numerosos profesionales de la lengua como lexicógrafos, traductores, profesores, etc. Para acceder a la plataforma, es necesario registrarse. Para tener acceso al corpus, hay que solicitarlo a la autora, indicando el correo electrónico, para poder darle de alta como usuario.

Con esta herramienta, es posible obtener varios tipos de resultados: concordancias (*Concordance*), listas de palabras (*Wordlist*), colocaciones y combinaciones de palabras (*Word Sketch*), sinónimos o palabras semejantes (*Thesaurus*), comparaciones entre las colocaciones de dos palabras (*Word Sketch Difference*) y expresiones multipalabra (*N-grams*).

En la Figura 1, pueden verse por ejemplo las líneas de concordancia resultado de la búsqueda de la colocación del verbo *ser* seguido por el adverbio *bien*, un error común entre los aprendices de español.

[15]Según la lista del *Plan Curricular* del Instituto Cervantes (2007).
[16]https://www.sketchengine.eu.

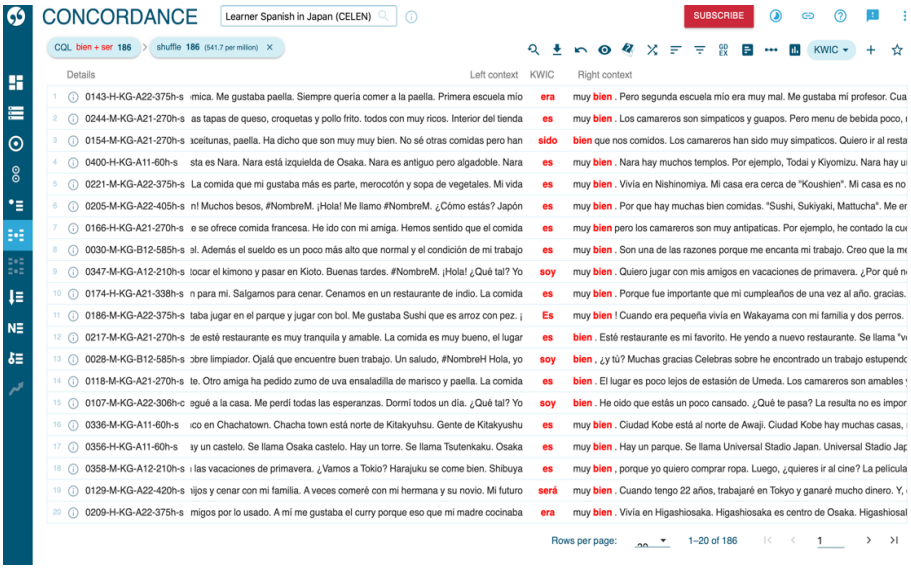

Figura 1. Líneas de concordancia del verbo *ser* seguido por el adverbio *bien*.

También es posible descargar el corpus completo y manipularlo directamente, bajo una licencia CC BY-NC 4.0. Véase la página web del proyecto: https://sites.google.com/view/celen/.

En el futuro, seguiremos ampliando el corpus con textos de otras instituciones, lo que permitiría, a partir de un muestreo de los datos, crear un corpus de referencia del español en Japón. También estudiamos la posibilidad de incluir textos procedentes de contextos de práctica informales como los blogs y las redes sociales para el aprendizaje de lenguas extranjeras y aumentar la precisión de la anotación morfosintáctica automática revisando los errores más frecuentes.

AGRADECIMIENTOS

Esta investigación ha sido posible gracias a todos los profesores de español de la Universidad Kansai Gaidai que colaboraron amable y desinteresadamente en la recogida de cuestionarios y textos y a los estudiantes que generosamente aceptaron participar cediendo sus datos.

La investigación ha sido financiada por *kakenhi* (17H07270), *Grant-in-Aid for Scientific Research* de la Japan Society for the Promotion of Science.

APÉNDICE

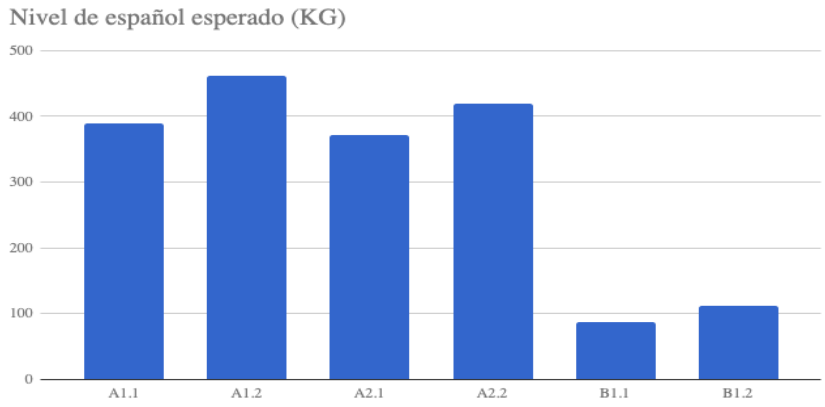

Figura 1. Número de textos según el nivel del MCER esperado

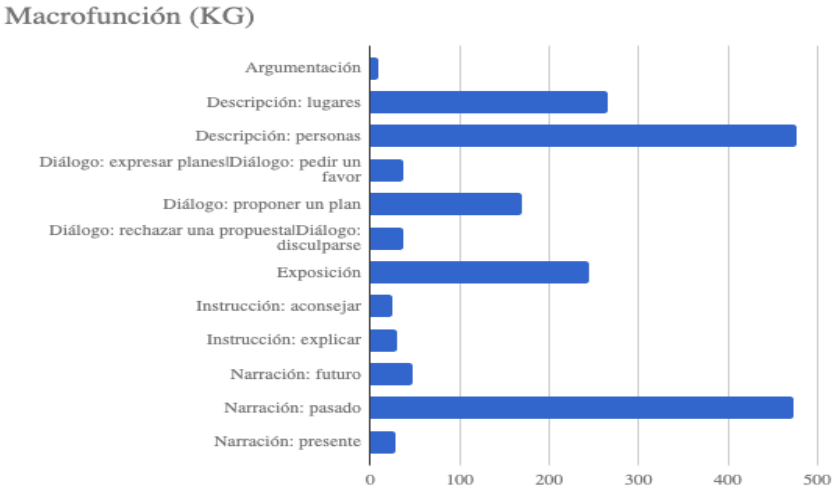

Figura 2. Número de textos según su macrofunción predominante

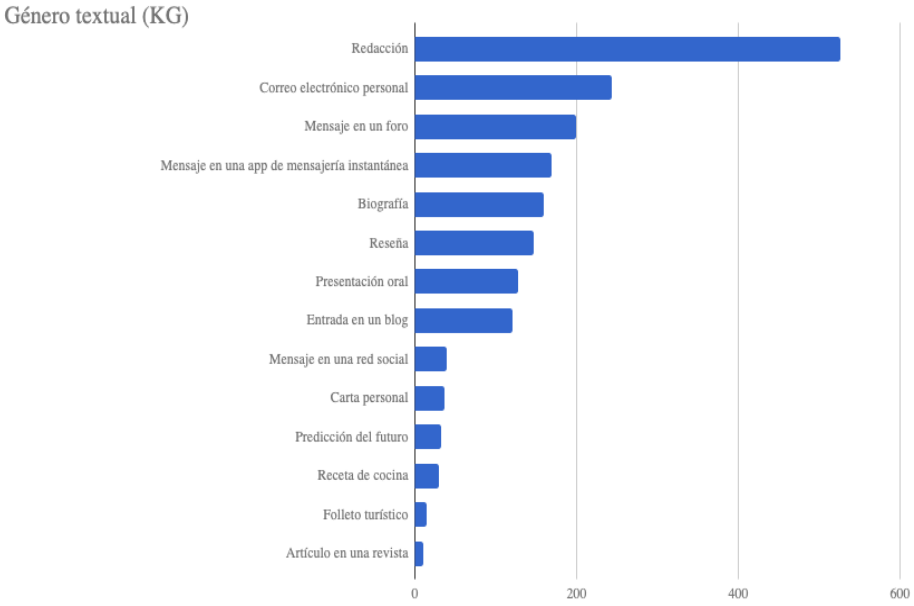

Figura 3. Número de textos según su género textual

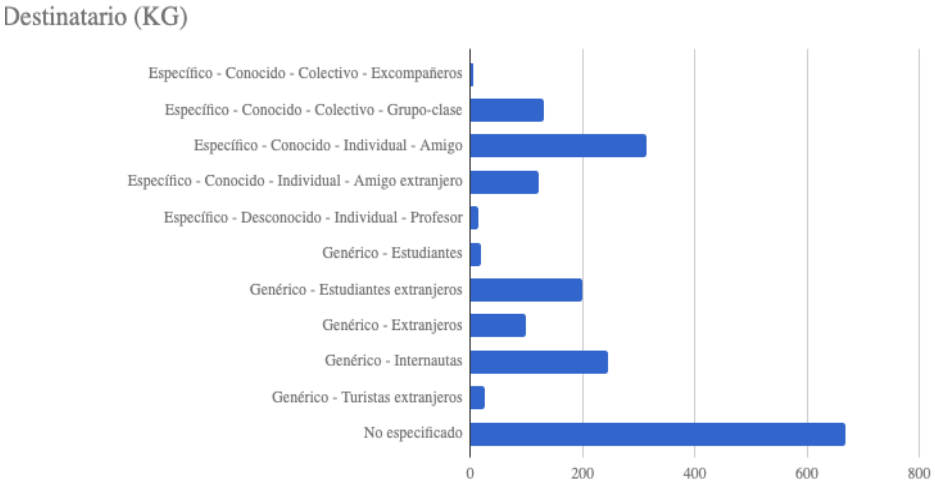

Figura 4. Número de textos según el tipo de destinatario

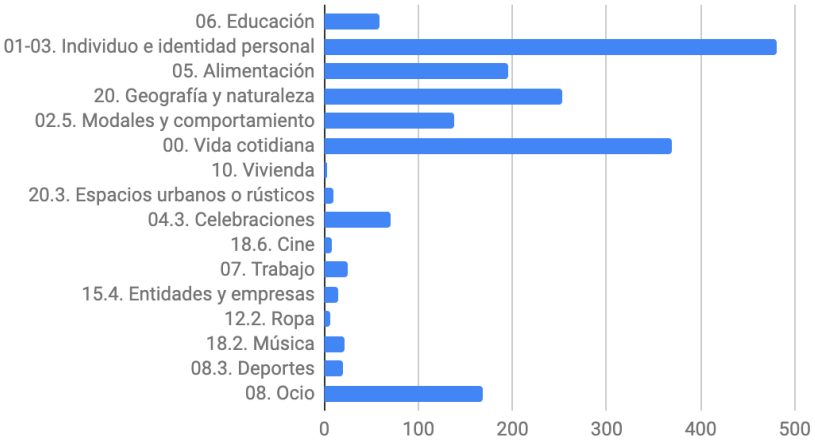

Figura 5. Número de textos según la noción específica predominante

REFERENCIAS BIBLIOGRÁFICAS

Alsonso-Ramos, M. 2016. Spanish Learner Corpus Research. Current trends and future perspectives. John Benjamins.

Berdicevskis, A. 2018. "Do non-native speakers create a pressure towards simplification? Corpus evidence". En Cuskley, C., Flaherty, M., Little, H., McCrohon, L., Ravignani, A. & Verhoef, T. (Eds.): *The Evolution of Language: Proceedings of the 12th International Conference (EVOLANGXII)*. 41-43.

Boyd A., Hana J., Nicolas L., Meurers D., Wisniewski K., Abel A., Schone K., Stinlova B. y Vettori C. 2014. "The MERLIN corpus: Learner language and the CEFR", *Proceedings of LREC 2014*. Reykjavik, Iceland: European Language Resources Association.

Buyse K. 2012. "El corpus de aprendices Aprescrilov y su utilidad para la didáctica de ELE en la Bélgica multilingüe", *Actas de ASELE 2012*.

Cestero Mancera A.M., Penadés Martínez I., Blanco Canales A., Camargo Fernández L. y Simón Granda J.F. 2001. "Corpus para el análisis de errores de aprendices de E/LE (CORANE)", *Actas de ASELE 2001*. 527– 534.

Escandón A. 2017. "La enseñanza universitaria de ELE/EL2 en Japón, tiempo de reformas", *Boletín de la Asociación para la Enseñanza del Español como Lengua Extranjera*, ASELE, 56. 25-29.

Gide. 2012. *Cuestionario sobre Análisis de Necesidades Aplicado a los Alumnos Universitarios Japoneses de Español. Informe*. GIDE, Grupo de Investigación de la Didáctica del Español. En línea: http://gide.curhost.com/publicacionesCuest.html.

Gide. 2015. *Un modelo de contenidos para un modelo de actuación. Enseñar español como segunda lengua extranjera en Japón*. GIDE, Grupo de investigación de la didáctica del español. En línea: http://gide.curhost.com/archivos/201601Modelo.pdf.

Granger S., Dagneaux E., Meunier F., Paquot M. 2009. *International Corpus of Learner English v2* (Handbook + CD-Rom). Presses universitaires de Louvain, Louvain-la-Neuve.

Instituto Cervantes. 2007. *Plan Curricular del Instituto Cervantes. Niveles de*

referencia para el español. Madrid: Biblioteca Nueva. En línea: https://cvc.cervantes.es/ENSENANZA/biblioteca_ele/pl an_curricular/indice.htm

Instituto Cervantes. 2018. *El español: una lengua viva 2018*. En línea: https://cvc.cervantes.es/lengua/espanol_lengua_viva/.

Llorián González S. 2017. "Claves de una revisión de los Niveles de Referencia para el Español, basada en metodología de corpus". *Marcoele Revista de Didáctica de ELE*, 25.

Lozano C. y Mendikoetxea A. 2013. "Learner corpora and second language acquisition: the design and collection of CEDEL2", *Automatic Treatment and Analysis of Learner Corpus Data,* N. Ballier, A. Díaz-Negrillo y & P. Thompson (eds.). Amsterdam: John Benjamins. 65–100.

Mizumoto T., Komachi M., Nagata M. y Matsumoto Y. 2011. "Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners", *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP),* Chiang Mai, Thailand.147-155.

Moreno García C. 2015. "Necesidades en la enseñanza del español en universidades japonesas. Reflexiones y sugerencias", *Actas del II Congreso Internacional sobre el español y la cultura hispánica del Instituto Cervantes de Tokio*. Tokio: Instituto Cervantes de Tokio. 66-77.

Mukherjee J. y Rohrbach J-M. 2006. "Rethinking applied corpus linguistics from a language-pedagogical perspective: New departures in learner corpus research", *Planning, Gluing and Painting Corpora: Inside the Applied Linguistics Workshop*, B. Kettemann y G. Marko (eds.). Frankfurt:Peter Lang. 205-232.

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

Rojo G. y Palacios Martínez I. 2016. Learner Spanish on computer, *Spanish Learner Corpus Research: Current trends and future perspectives*, M. Alonso (ed.), John Benjamins. 55-87.

Sanz Yagüe M. 2013. "Situación presente y futura del español en Japón. Elucubraciones de una profesora dirigidas a otros profesores en Asia (y a las autoridades)", Jornadas organizadas por la editorial Edinumen y

la Universidad de Hong Kong, con la colaboración de la Consejería de Educación en China y el Instituto Cervantes de Pekín.

Sanz Yagüe M., Escandón Godoy A., Romero Díaz J., Ramírez Gómez D. y Civit i Contra R. 2015. *Enseñar español en Japón. Algunos aspectos de la enseñanza a japoneses*. Kobe City University of Foreign Studies Annals of Foreign Studies, 89. En línea: http://id.nii.ac.jp/1085/00001825/.

Seidlhofer B. 2002. "Pedagogy and local learner corpora. Working with learning-driven data". *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, S. Granger, J. Hung y S. Petch-Tyson (eds.). Philadelphia: John Benjamins. 213–234.

Tono Y. 2003. "Learner corpora: Design, development and applications", *Proceedings of the 2003 Corpus Linguistics Conference* [ICREL Technical paper 16]. Lancaster University. 800-809.

Tono Y. 2016. "What is missing in learner corpus design?", *Spanish Learner Corpus Research: Current trends and future perspectives*, M. Alonso (ed.). John Benjamins. 33-52.

Ugarte V. 2012. "El español en Japón", *Anuario Cervantes 2012*. Instituto Cervantes. En línea: http://cvc.cervantes.es/lengua/anuario/anuario_12/ugarte/p01.htm.