TEISEL
Tecnologías para la investigación en segundas lenguas
ISSN 2696-676X
Universitat de Barcelona
Facultat de Filologia i Comunicació

El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE

Pilar VALVERDE
Universidad Kansai Gaidai (Japón)
pilar-vi@kansaigaidai.ac.jp
https://orcid.org/0000-0002-4208-9336

Resumen: En este artículo se presenta el Corpus de ELE en Japón, CELEN (https://ske.li/qqr), una colección de textos escritos por hablantes de japonés (L1) con distintos grados de dominio del español como lengua extranjera, desde el nivel A1 hasta el nivel C2 del MCER. Los datos proceden de (1) universidades en Japón, donde el español se estudia como asignatura de lengua extranjera o como carrera, y (2) contextos de interacción real en Internet, como blogs electrónicos y foros. La versión 1.2, de abril de 2023, consta de 6.196 textos escritos por 1.035 aprendices, con un total de 658.467 palabras. En el apartado 1 se resume brevemente la situación del español en Japón y los corpus de aprendices existentes. En el apartado 2 se describen las características principales de CELEN, el proceso de recogida y anotación de los datos y la interfaz de consulta. En el apartado 3 se ilustra su uso con varios tipos de búsquedas (concordancias, colocaciones, listas de palabras y n-gramas), aplicadas a fenómenos lingüísticos relevantes en la docencia o la investigación en ELE: el uso de *se*, las preposiciones, la concordancia de género, el orden de palabras, las colocaciones verbales, la frecuencia léxica o las secuencias de categorías gramaticales más frecuentes. Se trata de un recurso abierto, que se actualiza periódicamente, y esperamos que otros profesores e investigadores puedan albergar sus textos en él para ofrecer a la comunidad científica una amplia muestra de aprendices japoneses de español. En la página web del proyecto (https://sites.google.com/view/celen) se puede consultar la guía de uso detallada y descargar íntegramente algunas partes del corpus bajo una licencia CC BY-NC 4.0.

Palabras clave: español; lengua extranjera; corpus de aprendices; expresión escrita

Català:

El corpus d'aprenents japonesos CELEN y la seva aplicació a la docència i la investigació en ELE

***Resum**: En aquest article es presenta el corpus d'ELE al Japó, CELEN (https://ske.li/qqr), una col·lecció de textos escrits per parlants de japonès com a primera llengua amb diversos graus de domini de l'espanyol com a llengua estrangera, des del nivell A1 fins al nivell C2 del MCER. Les dades procedeixen de (1) universitats dels Japó, on l'espanyol s'estudia com a assignatura de llengua estrangera o com a carrera, y (2) àmbits d'interacció real a Internet com blogs electrònics y forums. La versió 1.2, d'abril de 2023, consta de 6.196 textos escrits per 1.035 aprenents, en total 658.467 paraules. A l'apartat 1 es resum breument la situació de l'espanyol al Japó i els corpus d'aprenents existents. A l'apartat 2 es descriuen les característiques principals de CELEN, el procés de recollida i anotació de les dades i la interfície de consulta. A l'apartat 3 se n'il·lustra el seu ús amb diversos tipus de cerques (concordances, col·locacions, llistes de paraules i n-grames) aplicades a alguns fenòmens lingüístics rellevants en la docència o la investigació de l'espanyol com a llengua estrangera: l'ús del clític* se, *les preposicions, la concordança de gènere, l'ordre de paraules, les col·locacions verbals, la freqüència lèxica o les seqüències de categories gramaticals més freqüents. Es tracta d'un recurs obert, que s'actualitza periòdicament, i esperem que d'altres professors i investigadors puguin albergar-hi els seus textos per oferir a la comunitat científica una àmplia mostra d'aprenents japonesos d'espanyol. A la pàgina web del projecte (https://sites.google.com/view/celen) es pot consultar la guia d'ús detallada i descarregar íntegrament algunes parts del corpus sota una llicència CC BY-NC 4.0.*

***Paraules clau**: espanyol, llengua estrangera; corpus d'aprenents; expressió escrita*

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

English:

The CELEN learner corpus and its application to teaching and research in Spanish as a foreign language

***Abstract**: This paper presents the CELEN corpus (https://ske.li/qqr), a collection of texts written by Japanese L1 speakers with different levels of proficiency in Spanish as a foreign language, from level A1 to level C2 of the CEFR. The data comes from (1) universities in Japan, where Spanish can be studied as a foreign language subject or as a major, and (2) contexts of real interaction on the Internet such as electronic blogs and forums. Version 1.2 (April 2023) is composed of 6,196 texts written by 1,035 learners, with a total of 658,467 words. In section 1 we briefly review the situation of Spanish as a foreign language in Japan and the existing learner corpora. In section 2 we describe the main features of the corpus, the data collection and annotation process and the search interface. In section 3 we exemplify various types of searches (concordances, collocations, word lists and n-grams) applied to linguistic phenomena relevant in the teaching and research of Spanish: the use of se, prepositions, gender agreement, word order, verbal collocations, lexical frequency, and pos-tag sequences. This is an open resource, that is updated periodically, and we hope that other teachers and researchers can include their texts in it and offer the scientific community a wide sample of texts from Japanese learners of Spanish. A detailed user guide is available on the project website (https://sites.google.com/view/celen) and parts of the corpus can be downloaded in full under a CC BY-NC 4.0 license.*

***Keywords**: Spanish; foreign language; learner corpus; writing*

Introducción

En los últimos años se han producido avances significativos en el campo de la investigación sobre corpus de aprendices de español (Alonso-Ramos, 2016). Este tipo de corpus, compuestos por textos producidos por personas que aprenden una lengua extranjera, son de gran utilidad para el procesamiento del lenguaje natural, la lexicografía pedagógica, la evaluación lingüística, la enseñanza de lenguas extranjeras y la investigación sobre la adquisición de segundas lenguas, entre otros (Callies y Paquot, 2015).

En el campo de la enseñanza del español como lengua extranjera (ELE), la formación en lingüística de corpus está cada vez más presente en los programas de formación del profesorado (Cruz, 2012; Elvira-García, 2021) y el interés por este tipo de recursos va en aumento. De hecho, los profesores de lenguas extranjeras emplean corpus de aprendices de manera informal constantemente: después de corregir las redacciones de un grupo de estudiantes, el profesor tiene claro cuáles han sido los errores más frecuentes en esa tarea y qué contenido tiene que repasar en la próxima sesión. Y aunque no llegue a crear un corpus propiamente dicho, es probable que, con el paso de los años, recopile una larga lista de dificultades y errores frecuentes, con ejemplos auténticos producidos por sus estudiantes. Sin embargo, todavía son escasas las publicaciones sobre el uso que los docentes hacen de estos recursos, en sus clases o en su propio desarrollo profesional.

CELEN (https://ske.li/qqr) es un corpus de aprendices desarrollado por y para profesores de ELE, para su aplicación en la formación de profesores, el diseño de materiales didácticos y actividades de aula, y la planificación curricular. En este ámbito, los corpus nos sirven para conocer la frecuencia de los elementos lingüísticos (por ejemplo, con qué frecuencia los aprendices usan ciertas palabras o estructuras) o responder dudas que no podemos resolver con nuestra propia intuición (por ejemplo, qué preposiciones suelen usar con ciertos verbos). Además de este uso pedagógico, para el que el corpus fue diseñado en un principio, los datos contenidos en él pueden ser de ayuda también en la investigación sobre el aprendizaje y la adquisición de lenguas extranjeras (en el análisis de errores, la lingüística

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

contrastiva o la lingüística cognitiva, entre otros), siempre teniendo en cuenta que no hay un corpus que sirva para todo (Hunston, 2008), sino que cada investigador debe evaluar cuidadosamente los recursos disponibles y elegir el más adecuado para su propósito. Por ejemplo, el subcorpus de blogs incluido en CELEN no contiene información sobre la edad de los participantes, aspecto que es necesario tener en cuenta si se pretende llevar a cabo una investigación en la que la edad del aprendiz sea relevante.

El objetivo del artículo es dar a conocer las características del corpus y ofrecer una pequeña muestra de sus múltiples posibilidades de explotación en el campo del ELE, con especial atención a la investigación. En el apartado 1 se resume brevemente la situación del español en Japón y los corpus de aprendices existentes. En el apartado 2 se describen las características principales de CELEN, el proceso de recogida y anotación de los datos y la interfaz de consulta. En el apartado 3 se describen varios tipos de búsquedas mediante ejemplos prácticos: el uso de *se*, las preposiciones, la concordancia de género, el orden de palabras, las colocaciones verbales, la frecuencia léxica o las secuencias de categorías gramaticales más frecuentes. Por último, se presentan las conclusiones y líneas de investigación futuras.

1. Contexto

En cuanto a la situación del español en Japón, en la encuesta realizada por el Instituto Cervantes de Tokio poco después de su apertura, en 2009, el español aparecía como la quinta lengua de interés entre los japoneses, por detrás del inglés, el coreano, el chino y el francés (Badillo, 2021), siendo las principales razones para estudiar español el interés por la cultura hispana y el hecho de que es una de las lenguas más habladas del mundo. Por niveles educativos, no es posible estudiar español en la enseñanza primaria (el inglés comienza a enseñarse en el tercer curso) y en la secundaria las cifras son extremadamente bajas. La mayoría de los aprendices de español está en la universidad, donde el español se puede estudiar en dos modalidades: como asignatura de lengua extranjera o como carrera. Aparte de la universidad, también es posible estudiar español fuera de la educación formal, en academias de idiomas, en el Instituto Cervantes de Tokio, en centros culturales, con profesores particulares o mediante los programas de radio y televisión de la cadena pública NHK. Para más detalles, véase Badillo (2021) y Moreno (2022).

Uno de los retos a la hora de recopilar datos para el corpus ha sido la relativa escasez de aprendices, especialmente de aquellos con un nivel intermedio o avanzado del idioma. Se estima que hay unos 60.000 estudiantes de español en Japón, ocupando el vigésimo puesto del ranking por países, liderado por Estados Unidos (más de ocho millones de estudiantes), Brasil (más de seis millones) y Francia (casi tres millones) (Instituto Cervantes, 2020:24). Por otra parte, se trata de una población muy homogénea en cuanto a conocimientos de lenguas: la primera lengua es casi siempre el japonés y el inglés suele ser la primera y única lengua extranjera.

En cuanto a los corpus de aprendices de español existentes (cf. Alonso-Ramos, 2016; Rojo *et al.*, 2023), si centramos nuestra atención exclusivamente en los corpus escritos, como el que aquí presentamos, y sin tener en cuenta

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

corpus de reducido tamaño compilados para llevar a cabo investigaciones particulares, existen en este momento tres que pueden ser consultados libremente (CEDEL2 (Lozano, 2022); CAES (Palacios *et al.*, 2019); y COWS-L2H (Yamada *et al.*, 2020)), otros tres mediante registro (Aprescrilov (Buyse y González, 2013), CATE (Lu, 2010) y CORESPI (Bailini y Frigerio, 2019)) y uno mediante compra (CORANE (Cestero Mancera *et al.*, 2001)).

En cuanto a datos de aprendices japoneses, solo disponemos de pequeñas muestras en tres corpus globales, que reúnen textos procedentes de diversas lenguas maternas (CORANE, CEDEL2 y CAES, en la Tabla 1). Los investigadores interesados en el componente oral tienen a su disposición dos corpus con datos de aprendices japoneses: Campillos Llanos (2014) y García Ruiz-Castillo (2022).

Tabla 1
Muestras de japonés L1-español L2 dentro de corpus globales, en orden cronológico. Elaboración propia.

Corpus	Palabras	Características de los textos	Procedencia de los datos
CORANE	55.000	Niveles A2-C1. Varios textos por aprendiz. Varios temas tratados en clase. 93% con herramientas de ayuda.	Universidad de Alcalá (España).
CEDEL2 (versión 2.0)	23.000	Niveles principiante - avanzado. Un texto por aprendiz. Elección entre 14 temas propuestos. 51% con herramientas de ayuda.	Principalmente universidades en Japón.
CAES (versión 2.1)	105.000	Niveles A1-C1. Varios textos por aprendiz. 13 temas en total (2 o 3 temas por nivel).	Principalmente Instituto Cervantes de Tokio y Universidad de Estudios Extranjeros de Kobe.

El corpus CORANE (Cestero Mancera *et al.*, 2001) fue uno de los primeros corpus de aprendices de español y participaron en él hablantes de distintas lenguas maternas, todos ellos alumnos de los Cursos de Lengua y Cultura Españolas para Extranjeros de la Universidad de Alcalá, en España. Los datos se recogieron en el año 2000 y fueron publicados en 2009 como CD-ROM. Como muestra la Tabla 1, en el subcorpus japonés el 93% de las composiciones fueron realizadas con acceso a herramientas de ayuda. En cuanto a los temas y tipos de texto, no se parte de una clasificación preestablecida sino que cada aprendiz contribuye con varios textos variados, la mayoría composiciones pero también ejercicios para practicar vocabulario o conversaciones.

CEDEL2 (http://cedel2.learnercorpora.com/) (Lozano, 2022) es un corpus de aprendices destinado a la investigación sobre la adquisición de segundas lenguas que recoge datos desde 2006 a través de un formulario en línea, donde cada voluntario escribe un texto sobre un tema elegido libremente entre catorce temas propuestos. En 2020 se incorporó un subcorpus de L1 japonés, como muestra la Tabla 1. El nivel de los textos se estima con varias medidas: el resultado de una prueba de nivel compuesta por 43 preguntas de opción múltiple, la autoevaluación del nivel por el propio participante en una escala del 1 al 6, el curso académico y la obtención de títulos oficiales de español.

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

Por último, el CAES (https://galvan.usc.es/caes) (Palacios *et al*. 2019), promovido por el Instituto Cervantes y desarrollado en la Universidad de Santiago de Compostela, recoge muestras en distintos centros del Instituto Cervantes y universidades desde 2011, mediante una aplicación informática y de acuerdo con unos criterios unitarios y un protocolo común de actuación. Los participantes escriben entre una y tres tareas escritas que varían en su grado de complejidad de acuerdo con el nivel acreditado por el estudiante, con trece temas en total. Es el único corpus de aprendices de español en el que los textos han sido desambiguados manualmente después de la anotación morfosintáctica automática.

Es evidente que las características de cada recurso vienen determinadas por el campo del que proceden los investigadores implicados o el uso que se le pretende dar. Así, podemos diferenciar tres corrientes principales o tipos de corpus: a) dirigidos a profesionales de la enseñanza b) enfocados a la investigación sobre adquisición de segundas lenguas y c) destinados al desarrollo de aplicaciones de procesamiento del lenguaje natural. CELEN se encuentra en la primera categoría, ya que ha sido desarrollado y está destinado principalmente a profesores de español y creadores de materiales, que van a usar el corpus principalmente con fines pedagógicos, como fuente de datos reales para conocer mejor la interlengua de los aprendices. No obstante, en su diseño se ha intentado favorecer su aprovechamiento por parte de investigadores de otros campos como la adquisición de segundas lenguas o la lingüística computacional. Para ello, se han registrado numerosas variables sobre el aprendiz y sobre la situación en la que se escribió cada texto, de manera que cada investigador puede seleccionar los textos que sean de interés para su investigación particular. Por ejemplo, es posible seleccionar solo los textos escritos por aprendices de nivel A2, en condiciones de examen (sin acceso a herramientas de ayuda), o por aprendices que viven en un determinado país.

2. El corpus CELEN

2.1. Características

CELEN es el primer corpus escrito disponible dedicado exclusivamente a aprendices nipohablantes y también el primero que incluye datos naturales extraídos de Internet. La versión 1.2, de abril de 2023, está compuesta por 6.196 textos escritos por 1.035 aprendices, con un total de 658.467 palabras (790.086 *tokens*).

El objetivo es ofrecer una muestra de los textos que los aprendices escriben durante su proceso de aprendizaje del español, en la universidad, o fuera de ella, en plataformas en línea. En el proceso de diseño y recogida, se han seguido las recomendaciones generales de la lingüística de corpus (Sinclair, 2005; McEnery y Hardie, 2011; Schäfer y Bildhauer, 2013; Parodi, 2022) y se ha puesto especial atención en la documentación y registro de toda la información disponible sobre las características del aprendiz y del texto[1]. Abordamos la construcción del corpus como un proceso

[1] La lista completa de metadatos puede consultarse en https://sites.google.com/view/celen/documentación/metadatos/ La distribución de frecuencia de cada metadato puede verse dentro de la aplicación de consulta, con la herramienta *Text type analysis*.

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

cíclico, en el que se identifican fortalezas y carencias y se añade o elimina material para corregirlas (Atkins *et al.*, 1992:4).

Concretamente, los textos proceden de varios contextos: a) el ámbito universitario en Japón (56% de *tokens*), donde el español puede estudiarse como asignatura de lengua extranjera o como carrera, y b) contextos de interacción real en Internet (44%), como blogs electrónicos y foros.

En cuanto a sus características principales (Gilquin, 2015), se trata de un corpus longitudinal, es decir, que recoge varias muestras de los mismos aprendices a lo largo de cierto periodo de tiempo. Actualmente, se dispone de datos longitudinales para el 80% de ellos y el periodo de seguimiento es de unos 15 meses de media (entre 7 y 9 meses para los participantes procedentes de universidades), con una horquilla que va desde un mínimo de un mes hasta más de 10 años. Los textos fueron escritos entre 2004 y 2022 e introducidos en el corpus entre 2015 y 2022.

Los participantes universitarios presentan un perfil bastante homogéneo: dos tercios son mujeres, su edad oscila entre los 18 y los 21 años, con un predominio del nivel básico de español, y muy pocos han estado más de un mes en países hispanohablantes. Los autores de textos publicados en Internet son más heterogéneos en cuanto a su edad, país de residencia y nivel de lengua, con predominio de los niveles B y C del MCER (*Marco Europeo de Referencia de las Lenguas*).

Se incluyen distintos géneros textuales (redacción, entrada en un blog, biografía, reseña, presentación, etc.) y una gran variedad de temas, desde los que se encuentran típicamente en los libros de texto usados en las aulas (describir a una persona, mi rutina diaria, mi infancia…) hasta temas más polémicos, actuales o privados, procedentes de interacciones reales en línea (la energía nuclear, el COVID-19, el nacimiento de un hijo…). Se indica si el aprendiz realizó la tarea con acceso a herramientas de ayuda -por ejemplo, si escribió el texto en casa, aunque se desconoce si finalmente utilizó dichas herramientas- o sin ellas.

2.2 Procedencia de los datos

Como se ha comentado en el apartado 1, uno de los retos a la hora de recopilar datos para el corpus fue la relativa escasez de aprendices, especialmente en los niveles intermedio y avanzado. Entre los estudiantes universitarios, cada año ingresan en la carrera de español apenas mil estudiantes nuevos en todo el país, en poco más de una decena de universidades. Para obtener la mayor representatividad y cantidad de datos posible, se optó por combinar datos de varias fuentes:

- El contexto universitario. Se recogieron datos en dos universidades que representan dos modalidades de estudio: la Universidad Kansai Gaidai, donde el español se estudia como carrera, y la Universidad de Kioto, donde se estudia como asignatura de segunda lengua extranjera de nivel inicial. Se trata de textos escritos como parte de un examen, tarea o actividad de clase.

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

- Contextos de aprendizaje informal en Internet. Se incluyen entradas publicadas en blogs electrónicos y en la red social para aprendizaje de lenguas extranjeras Lang-8, y mensajes publicados en el foro *Sólo español* de WordReference. Se trata de textos escritos por iniciativa propia: entradas de un blog o mensajes en un foro.
- Corpus no publicados: El *Japanese Learner Corpus of Spanish* (JALCOS), un pequeño corpus recopilado en 2004 en varias universidades que no había sido publicado hasta este momento. Los participantes escribieron un texto a propósito cada uno, para participar en la investigación.

CELEN es el primer corpus de aprendices que incluye datos naturales extraídos de Internet, lo que presenta varias ventajas. En primer lugar, la recopilación con métodos tradicionales es lenta y laboriosa, pero la web y el aumento de las plataformas de aprendizaje y evaluación en línea hacen posible acceder a grandes cantidades de textos de forma inmediata (Alexopoulou *et al.*, 2022). Con los corpus de aprendices de español existentes, que apenas superan el millón de palabras, es posible estudiar fenómenos gramaticales muy frecuentes (uso de artículos, preposiciones, alternancia de *ser* y *estar*…). Sin embargo, en el aula de lenguas extranjeras el vocabulario desempeña un papel tan importante como la gramática (Lewis, 1993 y 2000; Nation, 2006) y para su estudio es necesario desarrollar recursos de un tamaño mucho mayor que los actuales ya que, como es sabido, debido a la ley de Zipf, la mitad de las palabras de cualquier corpus aparece solo una vez, mientras que son necesarias al menos 20 ocurrencias de una palabra para describir su comportamiento (Sinclair, 2005).

En segundo lugar, los corpus web suelen contener una mayor variedad de temas y tipos de texto. El hecho de que estos no estén equilibrados, en el sentido de que no siguen un diseño preestablecido, no parece ser un gran inconveniente, ya que su gran tamaño y variedad lo compensa. Como señala Leech (2007: 138) "[…] in general, the larger a corpus is, and the more diverse it is in terms of genres and other language varieties, the more balanced and representative it will be." Del mismo modo que en el discurso nativo los distintos registros muestran distribuciones particulares de rasgos lingüísticos (Davies *et al.*, 2006), en el discurso de los aprendices los distintos tipos de tareas también influyen en la frecuencia de ciertos elementos lingüísticos (Brook y Hirst, 2013), en la complejidad del discurso y en su corrección (Tracy-Ventura y Myles, 2015). Por lo tanto, para representar mejor la interlengua es necesario incluir una gran variedad de tipos de texto y condiciones de producción. En cuanto a la representatividad, coincidimos con Kilgarriff y Greffenstette (2003: 343) en que "the web is not representative of anything else. But nor are other corpora, in any well-understood sense".

En tercer lugar, los textos que componen la mayoría de corpus (incluyendo CELEN) se recopilan en un entorno de enseñanza un tanto cerrado, en el aula o a propósito para un proyecto de investigación, y no en interacciones reales, y las tareas propuestas están relativamente alejadas de la vida cotidiana (Rojo et al, 2023:10). En cambio, los textos publicados en blogs o foros han sido escritos por iniciativa propia y se trata de muestras auténticas, que reflejan cómo escriben los alumnos cuando no están en el aula y qué temas les interesan.

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

En cuarto lugar, la escritura electrónica constituye un género propio, entre el discurso oral y el escrito, que merece ser representado en cualquier corpus. Uno de los cambios educativos más relevantes de la última década es el fácil acceso a múltiples productos culturales que facilitan el contacto con la lengua y la cultura españolas desde cualquier parte del mundo (Cassany, 2023) y la comunicación digital o "comunicación real mediada por ordenador", que representa una parte importante de la interacción humana, adquiere especial relevancia en contextos donde las oportunidades de comunicación auténtica en la lengua extranjera son escasas, como es el caso del español en Japón. Como añadidura, los blogs o diarios electrónicos se actualizan con frecuencia, siendo posible estudiar la evolución de un mismo autor a través de un cierto periodo de tiempo.

Como contrapartida, la extracción de datos de la web presenta algunos inconvenientes. En primer lugar, los textos suelen contener más "ruido" (fragmentos en otras lenguas, documentos demasiado cortos o largos, inclusión de material ajeno como letras de canciones o noticias), que debe ser eliminado de forma totalmente automática con cierto margen de error, ya que es inviable inspeccionar a mano miles de documentos. En segundo lugar, la productividad de cada aprendiz no está equilibrada, sino que unos pocos autores (generalmente los de niveles más altos) escriben mucho y la gran mayoría escriben poco. En tercer lugar, la información sobre el perfil del aprendiz o las condiciones en las que se produjo el texto es limitada (Davies, 2013). Por último, la distribución de textos procedentes de la web es problemática: está permitido descargar líneas de concordancia, pero no redistribuir los textos íntegros, sujetos a copyright[2].

El contexto universitario

En la Universidad Kansai Gaidai, donde el español se estudia como carrera -y donde se encuentran una cuarta parte de los estudiantes de esta especialidad del país- se repartió un cuestionario entre los estudiantes al principio del año académico, donde se solicitaron datos personales básicos, variables sociolingüísticas y de experiencia lingüística, así como el consentimiento informado. Durante el curso, se recogieron composiciones escritas como parte de exámenes, en las clases impartidas por profesores nativos (en primer y segundo curso), así como en tareas o actividades de clase (en tercer curso). Los datos se recogieron en papel y fueron transcritos mediante la herramienta de OCR de *Google Drive* y posterior revisión manual. El nivel asignado es el curso académico y el nivel del MCER del libro de texto utilizado en el aula, entre A1 y B1[3]. El subcorpus resultante contiene unas 141.000 palabras, en 1.840 textos de 459 alumnos (unos cuatro textos por estudiante, escritos a lo largo de un año académico). Los títulos de las tareas (31 en

[2] Las leyes sobre copyright varían en cada país y los métodos para reducir estas limitaciones son variados. Algunos autores como Mark Davies, por ejemplo, optan por transformar los textos, reemplazando un fragmento de diez palabras por una secuencia de símbolos, cada 200 palabras (https://www.corpusdata.org/limitations.asp), mientras que para ciertas aplicaciones de PLN puede ser suficiente desordenar las oraciones del corpus, de modo que no sea posible identificar el origen de los datos.

[3] Los aprendices empiezan sus estudios sin conocimientos previos de español, y es necesario aprobar la clase anterior para entrar en las clases superiores, lo que garantiza cierta homogeneidad en los niveles de dominio dentro de cada clase.

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

total) son, entre otros: *Una presentación personal, Costumbres sociales de Japón, Una persona que admiro, Mi plan para las vacaciones, Un lugar conocido o que me gusta, Mi infancia, Mi visita a un restaurante español, Un día importante en mi vida, Descripción de un accidente geográfico del mundo, Un mal día*, etc. Para una descripción más detallada del proceso de recogida, véase Valverde (2020).

En la Universidad de Kioto, donde estudiantes de distintas facultades pueden cursar una clase de español inicial de un año de duración, el proceso de recogida de los datos fue menos laborioso. Los estudiantes de esta universidad, más familiarizados con la tecnología, presentan todas sus tareas escritas por medios electrónicos, y se les anima a aprender a utilizar herramientas de ayuda a la escritura como diccionario, corrector ortográfico, etc. Al final del curso se pidió a los estudiantes que rellenaran un breve cuestionario (edad y conocimiento de otras lenguas) y esta información se añadió a las composiciones que habían entregado durante el curso, manteniendo el anonimato durante todo el proceso. A estos textos les asignamos el nivel A1, ya que los alumnos son completamente principiantes en el idioma, aunque el objetivo de la clase es dar una visión completa del sistema lingüístico en un año, y los temas gramaticales que se practican en clase, especialmente en el segundo semestre, corresponden a niveles superiores de dominio. El subcorpus resultante contiene unas 144.000 palabras, en 2.111 textos de 278 alumnos (unos ocho textos por estudiante, escritos a lo largo de un año académico). Los títulos de las tareas (ocho en total) son, en orden cronológico: *Describir a una persona, Describir una ciudad, Describir una comida, La ropa que llevo hoy, Mi rutina diaria, Consejos para estudiar inglés, Cómo tirar la basura doméstica correctamente* y *Comentar una estadística sobre los residentes extranjeros en Japón*.

Contextos de aprendizaje informal en Internet: blogs electrónicos y foros

Para la construcción del subcorpus web, se realizó una búsqueda de blogs escritos en español, por hablantes de japonés, principalmente en los dominios de Blogger (https://www.blogger.com/) y WordPress (https://wordpress.com/es/). Se descargaron todas las entradas de cada blog, se sometieron a varias etapas de post-procesamiento y se asignaron algunos metadatos como el sexo y la ubicación, a partir de la información publicada en el perfil y en las entradas[4]. Para una descripción más detallada del proceso de creación de este subcorpus, véase Valverde (2018) y Schäfer y Bildhauer (2013). La primera versión, recopilada en 2015, tenía un tamaño de 625.000 palabras, y fue ampliada en 2022 con las nuevas entradas de los blogs que seguían en activo, obteniendo 880.000 en total. A pesar de su gran tamaño, era necesario abordar un inconveniente típico de los corpus web: hay unos pocos autores que escriben mucho (el usuario más activo produjo el 37% de los *tokens*), y muchos que escriben muy poco. Optamos por incluir en CELEN una muestra de entre 500 y 5.000 palabras por autor aproximadamente[5] y asignar manualmente un

[4] Además de la información publicada en el perfil, se confirmó si la información era correcta consultando las primeras entradas de cada blog, donde suele ser evidente si el autor del blog es un hombre o una mujer, y en qué país se encuentra.

[5] Se descartaron los autores que escribieron menos de 500 palabras en total. Se incluyeron todos los textos de los que escribieron entre 500 y 5.000 palabras. Para los que escribieron más de 5.000, se extrajo una muestra de 5.000 palabras, compuesta por textos completos. Por ejemplo,

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

nivel amplio (A, B, C) a cada uno, ya que no sería posible asignar de forma fiable un nivel más específico a una serie de textos escritos durante un periodo de tiempo tan largo. El subcorpus resultante contiene unas 119.000 palabras, en 556 textos de 29 aprendices que publicaron entradas en sus blogs entre 2004 y 2022 (una media de veinte textos por autor). Cada texto o entrada de blog trata sobre un tema distinto, elegido por el propio autor, cuyo objetivo principal es a) documentar su vida diaria en Japón u otro país de residencia, a modo de diario personal (*La boda de mi amiga, Cuando hace frío, Con mis amigos*…), b) explicar temas relacionados con la cultura japonesa a los internautas de otros países (*Día de los Niños, El cerezo, El water inteligente, Los palillos*…), a menudo recetas de cocina o comentarios gastronómicos (*La sopa de miso, El bento, Kinpira de raíz de loto, Bavarois de Matcha*…), c) explicar la lengua japonesa (*Hiragana, Frases de moda en 2013*…) o española (*Estar al quite, A renglón seguido*…) o d) discutir sobre temas de actualidad (*Una manifestación en Shibuya hoy, Un resumen del gobierno japonés*…).

En segundo lugar, para ampliar el corpus web anterior, añadimos algunos datos del corpus NAIST Lang-8 (Mizumoto *et al.*, 2011). Este es un gran conjunto de datos extraído de la red social Lang-8 (https://lang-8.com/), dedicada al intercambio de idiomas, y usado principalmente para el entrenamiento de sistemas de corrección automática de errores. Extrajimos los datos correspondientes a hablantes de japonés que aprenden español (solo el 1% de los usuarios), obteniendo unas 479.000 palabras. Al igual que en el corpus anterior, seleccionamos una muestra de entre 2.000 y 5.000 palabras por usuario, y asignamos manualmente un nivel amplio a cada uno. El subcorpus resultante contiene unas 149.000 palabras, en 1.247 textos de 39 aprendices (una media de 32 textos por autor) que participaron en la red social entre 2008 y 2011. Como en el caso anterior, cada texto trata sobre un tema distinto, elegido por el propio autor, con predominio de la narración de la vida cotidiana (*¡Hola!, Mi hijo, Mi habitación, Un viaje, El cumpleaños, Copa América, Ayer fui a ver flamenco, Dolor de cabeza*…).

Por último, incluimos también datos del corpus WordReference (Berdicevskis, 2020), que contiene los mensajes publicados en los foros de lengua del mismo nombre, por aprendices de cuatro idiomas (inglés, español, francés e italiano). Seleccionamos los mensajes de los hablantes de japonés como L1 publicados en el foro *Sólo español* (https://forum.wordreference.com/forums/sólo-español.45/), unas 50.000 palabras en total, e incorporamos a CELEN una muestra de hasta 5.000 palabras por autor. Asignamos el nivel C2 a todos ellos, ya que se trata de usuarios muy competentes o profesionales de la lengua que formulan preguntas sobre el español, de nivel avanzado. El subcorpus resultante contiene unas 19.000 palabras, en 220 textos de 8 usuarios (una media de 28 mensajes por autor), que participaron en el foro entre 2008 y 2019.

si un autor escribió 10.000 palabras, a partir de los textos en orden cronológico se seleccionaron textos a intervalos regulares para mantener la representatividad (en lugar de tomar solo la primera mitad o solo la segunda mitad) hasta llegar a la extensión total deseada, de 5.000 palabras.

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

Aprovechamiento de corpus no publicados

Durante la construcción de CELEN encontramos algunas referencias imprecisas a un recurso que parecía no estar disponible, el *Japanese Learner Corpus of Spanish* (JALCOS). Los datos de este proyecto, inspirado en el ICLE (*International Corpus of Learner English*), fueron recogidos en 2004 en cinco universidades de Japón, pero no llegaron a publicarse ni constituirse en forma de corpus, y nos fueron cedidos amablemente por su autor para incluirlos en CELEN. Los participantes, estudiantes de entre primer y cuarto curso de la carrera de español, escribieron un texto sobre un tema a elegir entre siete temas propuestos, y rellenaron un formulario de consentimiento y un cuestionario. Los temas son, de mayor a menor frecuencia: *¿Te gusta viajar?, ¿Qué opinas sobre el aprendizaje de lenguas extranjeras?, ¿Con qué huirías si ocurriera un gran terremoto?, Mi peor sueño, ¿Estudiar en la universidad es útil en el mundo real?, El mundo de los sueños y la imaginación en la sociedad moderna, ¿El dinero es la raíz de todos los males?*.

Para su inclusión en CELEN fue necesario dar forma a los datos conservados en formato electrónico hasta el momento: se extrajeron los metadatos de los cuestionarios administrados a los estudiantes, y se uniformizaron e incorporaron a la cabecera de cada texto. En cuanto al nivel de dominio, tomamos el curso académico del estudiante como referencia, y asignamos un nivel del MCER aproximado, basado en la siguiente correspondencia[6]: A1 (primer año), A2 (segundo año), B1 (tercer año), B2 (cuarto año). El subcorpus resultante contiene 87.000 palabras, en 222 textos de 222 alumnos, y está disponible para descarga también de forma independiente (https://doi.org/10.5281/zenodo.7768882).

2.3 Anotación

Los textos han pasado por varias etapas de procesamiento: anonimización, anotación no lingüística y anotación lingüística. En la cabecera de cada documento se incluye toda la información disponible sobre el aprendiz y el texto. La base de datos contiene 39 campos en total (24 sobre el aprendiz y 15 sobre el texto)[7], pero la cantidad de información varía en función del subcorpus. Por ejemplo, en los datos procedentes de la Universidad de Kioto no consta el sexo del aprendiz y en el subcorpus web se dispone de menos información, ya sea porque se desconoce ese metadato (por ejemplo, el nivel de otras lenguas extranjeras) o porque sería muy costoso introducirlo de forma manual (como la macrofunción textual o la noción específica). En el contexto educativo en el que se desarrolló CELEN consideramos como datos fundamentales, presentes en todos los textos: el nivel de dominio del MCER, el curso, la procedencia del texto (producción semi-espontánea en el aula o producción auténtica en Internet) y el acceso a herramientas de ayuda.

[6] Se trata solo de una aproximación, ya que desconocemos el nivel de español impartido en las universidades participantes en el año 2004, antes de la publicación de los niveles del MCER y del *Plan Curricular* del Instituto Cervantes. Actualmente el primer año de estudios suele equivaler al nivel A1 y el segundo curso a A2 o B1.1. El nivel en tercero y cuarto es más variable ya que depende del número de clases y de las estancias realizadas en el extranjero.

[7] La lista completa de metadatos puede consultarse en https://sites.google.com/view/celen/documentación/metadatos/ La distribución de frecuencia de cada metadato puede verse dentro de la aplicación de consulta, con la herramienta *Text type analysis*.

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

Como parte de la anotación no lingüística, se han introducido marcas que señalan el inicio y final de oración, párrafo y documento.

Por último, los textos han sido etiquetados automáticamente con el lema, la categoría gramatical y los rasgos morfosintácticos de cada palabra, mediante el analizador Freeling (https://nlp.lsi.upc.edu/freeling/) (Padró *et al.*, 2012). En esta versión del corpus la etiquetación es totalmente automática, no hay revisión manual. Por lo tanto, es necesario inspeccionar los resultados cuidadosamente y valorar si el margen de error es aceptable o no en cada caso particular. También es posible filtrar los resultados de manera semiautomática o ignorar la anotación y realizar búsquedas en el texto plano, sin anotar. Mientras que en textos ortográficamente correctos la precisión del etiquetador automático ronda el 97% (es decir, 3 de cada 100 palabras reciben una etiqueta incorrecta), en textos de aprendices de nivel intermedio la precisión baja hasta el 93% (Valverde, 2011), y previsiblemente más en el nivel inicial. A pesar del ruido generado por los errores del etiquetador, incluimos este nivel de análisis porque amplía las posibilidades de búsqueda.

2.4 Interfaz de consulta

Para dar acceso al corpus a la comunidad docente e investigadora elegimos la aplicación de consulta *Sketch Engine* (https://www.sketchengine.eu/), en su versión de acceso abierto. Se trata de un software para consulta y análisis de textos con una larga trayectoria, que goza de gran prestigio y popularidad entre profesionales de la lengua, y que ofrece no solo concordancias sino también herramientas para extraer colocaciones, listas, palabras semejantes, etc. Es de fácil manejo y a la vez permite realizar búsquedas avanzadas muy potentes mediante el lenguaje CQL. Para facilitar su aprendizaje, dispone de una guía de ayuda en la web, vídeos explicativos, así como cursos en línea y talleres presenciales para usuarios avanzados. Una gran ventaja de este sistema es que con la misma interfaz es posible consultar una gran cantidad de corpus en múltiples lenguas, rentabilizando el tiempo invertido en su aprendizaje y evitando tener que aprender a usar una herramienta nueva para cada corpus particular.

Además de la versión de pago, dispone también de una colección de corpus de consulta gratuita y sin registro (https://app.sketchengine.eu/#open), en la que se encuentra CELEN. Hemos creado también una página web (https://sites.google.com/view/celen) donde se describen las características principales del proyecto y se muestra paso a paso cómo usar cada herramienta mediante ejemplos prácticos.

En el tablero principal de la interfaz (Figura 1) se muestran las herramientas disponibles, de acuerdo con el tipo de resultado que se quiera obtener: concordancias, colocaciones, listas de palabras, tesauro, n-gramas, palabras clave, diferencias entre dos palabras y estadísticas de todo el corpus. El idioma por defecto es el inglés, pero se puede cambiar al español con la opción *Settings*, en la esquina superior derecha.

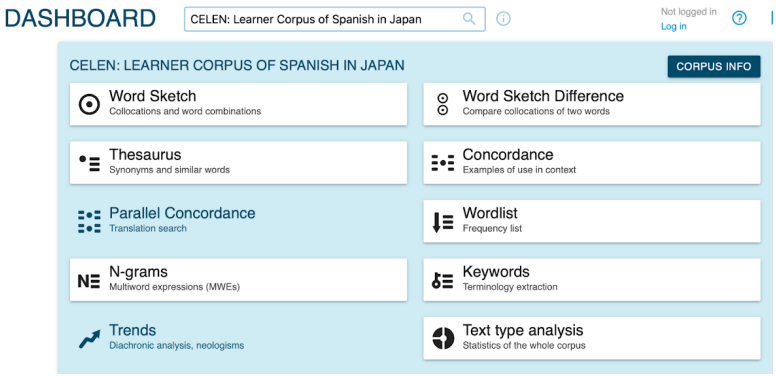
Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

Figura 1

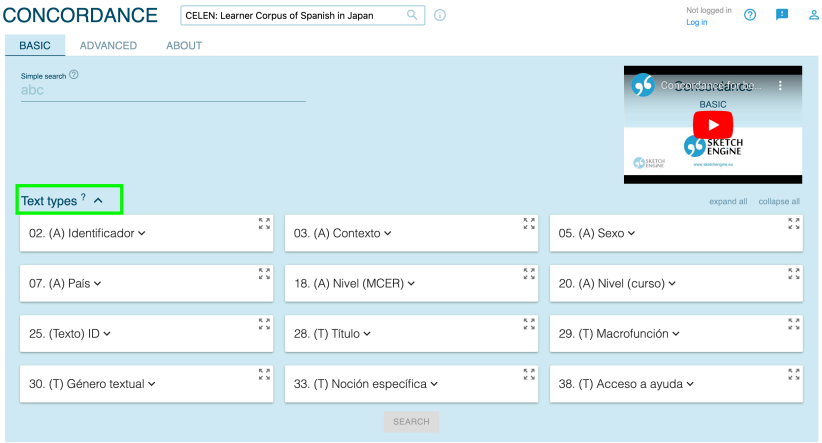
Tablero de la interfaz de consulta Sketch Engine. Herramientas de búsqueda



Las búsquedas pueden realizarse sobre todo el corpus o sobre alguna parte de este, de acuerdo con las características del aprendiz o del texto. Para ello, solo hay ir a cualquier herramienta (por ejemplo, *Concordancias*) y desplegar el menú *Text types*, como se muestra en la Figura 2.

Figura 2

Variables (Text types) sobre las que se pueden realizar las búsquedas



De los 39 metadatos que contiene la base de datos, se muestran solo 12, aquellos que nos han parecido de mayor interés general[8]. Para facilitar la lectura, se mantiene el número que identifica cada campo en la base de datos (02, 03, 05…). Los campos precedidos por (A) se refieren a las características del aprendiz, y los precedidos por (T), a las características del texto. Nótese que en el campo 18, *(A) Nivel (MCER),* (Figura 3), los niveles de dominio se muestran agrupados de forma jerárquica. De esta manera es posible seleccionar distintos grados de especificidad: solo el nivel A1, o todo el nivel A, por ejemplo.

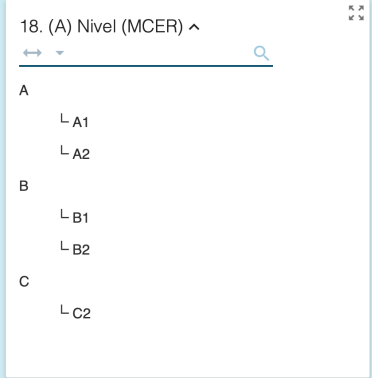
[8] Si se desea realizar las búsquedas sobre algún metadato que no se muestra en el menú *Text types*, es posible usar la búsqueda con CQL o ponerse en contacto con la autora para activarlo en la interfaz gráfica.

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

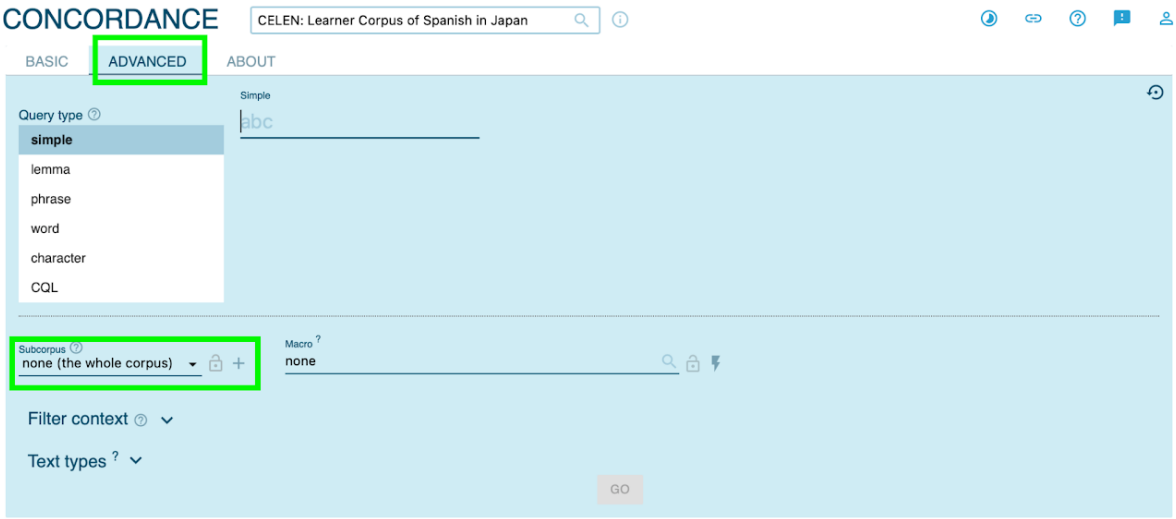
ISSN 2696-676X

Figura 3
Niveles de dominio (MCER)



Se ha optado por esta representación para reflejar más claramente los distintos niveles disponibles en cada subcorpus: mientras que en algunos se asigna un nivel específico a cada texto (A1, A2, B1…), en otros se asignan solo niveles amplios (A, B, C). Por lo tanto, todos los textos de nivel A1 o A2 pertenecen también al nivel A, pero no todos los de nivel A pertenecen a un subnivel como A1 o A2, sino que algunos de ellos están infraespecificados. Otra forma de delimitar las búsquedas es con la opción *Subcorpus*, en la pestaña de búsqueda avanzada (Figura 4).

Figura 4
Búsqueda sobre una parte del corpus (subcorpus)



Hay que tener en cuenta que los distintos subcorpus no han sido diseñados para ser comparados de manera independiente entre sí, sino que se trata simplemente de filtros. Cada investigador puede seleccionar los datos que son de su interés y valorar cómo su composición afecta a los resultados (véase por ejemplo Brooke y Hirst, 2013).

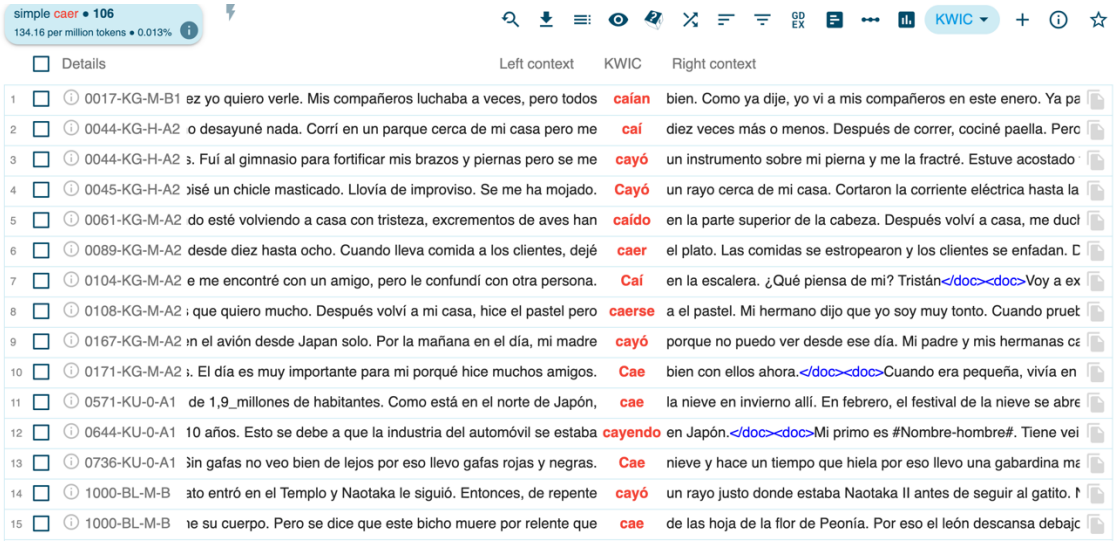
Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

El uso más típico del corpus consiste en buscar ejemplos de uso de una palabra. Para ello, simplemente hay que seleccionar la herramienta *Concordancias* y escribir la palabra deseada en la caja de búsqueda. Al introducir un lema cualquiera, se devuelve por defecto cualquiera de sus formas flexionadas, en minúsculas o mayúsculas, ya que este suele ser el tipo de resultado preferido por los usuarios. Si se escribe una forma flexionada, en cambio, se obtendrán solo los ejemplos de esa forma concreta. Si escribimos el verbo *caer* en infinitivo, por ejemplo, obtenemos el resultado que se muestra en la Figura 5: hay 106 casos, con o sin el clítico *se*, y en varios tiempos y modos[9].

Figura 5
Concordancias del verbo caer *(106 ocurrencias)*



A la izquierda de cada línea figura un código (*0017-KG-M-B1*) que contiene: el número de aprendiz (*0017*), un acrónimo que indica de dónde proceden los datos (*KG*)[10], el sexo (*M*) y el nivel del MCER (*B1*), lo que permite ver rápidamente cómo se distribuyen los resultados. También se pueden consultar todos los metadatos correspondientes a un ejemplo determinado, haciendo clic en el icono de información (i) que se encuentra a la izquierda de cada línea.

Los textos se presentan agrupados por aprendiz (en la imagen, 0017, 0044, 0045, etc.) y dentro de cada uno de ellos, en orden cronológico, lo que permite ver la evolución a la largo del tiempo y realizar estudios longitudinales.

En la barra de herramientas superior existen numerosas opciones de visualización y opciones para filtrar y resumir los resultados. Los botones imprescindibles son:

- El primer botón, *Change criteria*, para modificar la búsqueda actual, sin tener que volver al menú principal.

[9] Para obtener solo los ejemplos de la forma exacta *caer*, en infinitivo, hay que usar la opción de búsqueda avanzada *word*. Para obtener una cadena de palabras exacta (por ejemplo, *caer bien*, en infinitivo) hay que seleccionar la opción *phrase*.

[10] En orden alfabético: AP = Universidad Provincial de Aichi, BL = Blogger y WordPress, JSA = Universidad Sofía, KG = Universidad Kansai Gaidai, KU = Universidad de Kioto, L8 = Red social Lang-8, NZ = Universidad Nanzan, OF = Universidad de Estudios Extranjeros de Osaka, TH = Universidad de Tokoha, WR = WordReference.

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

- El segundo botón, *Download*, para descargar los resultados en varios formatos, con un contexto de hasta 100 caracteres a derecha e izquierda[11].
- El botón *View,* para mostrar las marcas que indican inicio y final de oración (<s>), párrafo (<p>) y documento (<doc>). Se recomienda mostrar la marca <doc> para ver claramente dónde empieza y dónde acaba cada texto (documento).
- El botón *Shuffle*, para ordenar las líneas aleatoriamente. Por defecto, las líneas se muestran en el orden en el que se incorporaron los textos (estos se agrupan por universidad, curso, clase, estudiante, etc.), pero, cuando hay cientos o miles de ocurrencias, puede ser preferible desordenarlas aleatoriamente para tener una visión global en la primera página de resultados.
- El botón *Frecuencia*, para obtener la distribución de frecuencia de la expresión buscada en función de las características del aprendiz o del texto, o para extraer la lista de formas o lemas.

3. Ejemplos prácticos

A nivel lingüístico, el español y el japonés son lenguas muy diferentes (véase por ejemplo Saito, 2005; Fukushima, 2014; Sanz *et al*., 2015; Takagaki, 2018). A continuación, a partir de unos ejemplos prácticos, se ofrece una muestra de las múltiples posibilidades de explotación de CELEN en el campo del ELE, con especial atención a la investigación: el uso de *se* (§ 3.1), las preposiciones, la concordancia de género, el orden de palabras (§ 3.2), las colocaciones verbales (§ 3.3), la frecuencia léxica (§ 3.4) y las secuencias de categorías gramaticales (§ 3.5).

3.1 Herramienta *Concordancias*: búsqueda simple

Los aprendices japoneses, al igual que los hablantes de otras lenguas maternas, tienen dificultades en el uso de los clíticos, especialmente los distintos **valores de *se***. Un caso típico es el verbo *caer(se),* con el sentido de "Moverse algo o alguien desde arriba hacia abajo por su propio peso". Se trata de un verbo cuyo aprendizaje plantea muchas dificultades por presentar una estructura muy compleja, ya que admite no solo el clítico de voz media para expresar involuntariedad (*caer/caerse*) sino también un dativo de interés (*caérsele algo a alguien*), y su estructura argumental está formada por un móvil que actúa como sujeto y un beneficiario animado que actúa como objeto indirecto, en lugar de la estructura prototípica sujeto animado-objeto inanimado.

Para buscar ejemplos de uso de este verbo, basta con ir a la herramienta de concordancias y escribir el lema del verbo -el infinitivo *caer,* sin clítico- en la casilla de búsqueda simple, como se ha mostrado en la Figura 5. Si

[11] Para descargar un contexto más amplio (hasta 500 caracteres) es necesario disponer de una cuenta en *Sketch Engine*, y acceder al corpus mediante la función "corpus compartido" (*shared corpus*). Para darle acceso, póngase en contacto con la autora indicando la dirección de correo electrónico de su cuenta.

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

inspeccionamos las líneas más de cerca, vemos que cuando este verbo va acompañado de un sujeto inanimado, sobre todo sin *se* (*cae la nieve, un rayo, las hojas, el sol, la lluvia, la tarde*), su uso es mayoritariamente correcto.

En su uso pronominal, sin embargo, encontramos algunos usos correctos con el clítico de voz media (b) pero también errores de todo tipo: con omisión del clítico de voz media (c), con omisión del dativo de interés (d), con un clítico de voz media superfluo (e), con un objeto directo en lugar de un sujeto (f) y con un sujeto (o tema) en lugar de un objeto indirecto (g). Estos resultados podrían ser un buen punto de partida para un análisis más profundo de las causas subyacentes a estos errores, relacionadas con la lengua materna de los aprendices, en que la idea de *caer* se expresa con el par de verbos intransitivo/transitivo 落ちる y 落とす(*ochiru* y *otosu*). En japonés, algo o alguien cae o se cae (con el verbo intransitivo *ochiru*) o alguien deja caer o pierde algo involuntariamente (con el verbo transitivo *otosu*). Esto explicaría los errores de los ejemplos (f) y (g), o la vacilación en el uso de los clíticos para expresar una involuntariedad que en parte es inherente al sentido de *caer* (c, e).

(a) Cayó un rayo cerca de mi casa (0045-KG-H-A2)

(b) Todavía no me he caido este año, pero siempre me caigo unas veces cuando saco a mis perras de paseo. (1042-L8-M-B)

(c) Nosotros huimos del hotel, es que algunas paredes cayeron y los cristales estaba roto. (0916-TH-M-B2)

(d) Sentí como realidad…"¡Dios mío, mis dientes… se cayeron !" (0808-JSA-0-A2)

(e) Hace unos días, mi hija se acostó sin bañarse ni ducharse porque había jugado demasiado durante el día y estaba tan cansada esa tarde que se cayó dormida. (1045-L8-H-B)

(f) Después volví a mi casa, hice el pastel pero caerse a el pastel. (0108-KG-M-A2)

(g) Una bailalína se le cayó un sombrero de juncia al suelo cuando bailaba con unas diez personas. (1024-L8-M-B)

3.2 Herramienta *Concordancias*: búsqueda avanzada con CQL

Para búsquedas más avanzadas, es recomendable usar CQL (*Corpus Query Language*), que es el lenguaje empleado internamente por la aplicación. Se pueden especificar condiciones de búsqueda en la etiqueta (*tag*), la forma de la palabra (*word*), el lema (*lemma*) o cualquier otro atributo. También se pueden expresar condiciones globales, para que la búsqueda tenga lugar dentro de los límites de la oración, el párrafo o el documento.

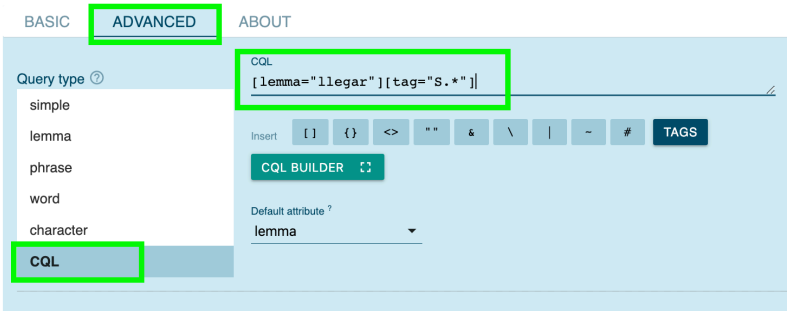
Una fuente de errores común entre los aprendices de cualquier lengua es el uso de las **preposiciones**. Para averiguar, por ejemplo, con qué preposiciones usan el verbo *llegar* los aprendices japoneses, se puede escribir la siguiente expresión en el menú de búsqueda avanzada, como en la Figura 6: `[lemma="llegar"][tag="S.*"]`

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

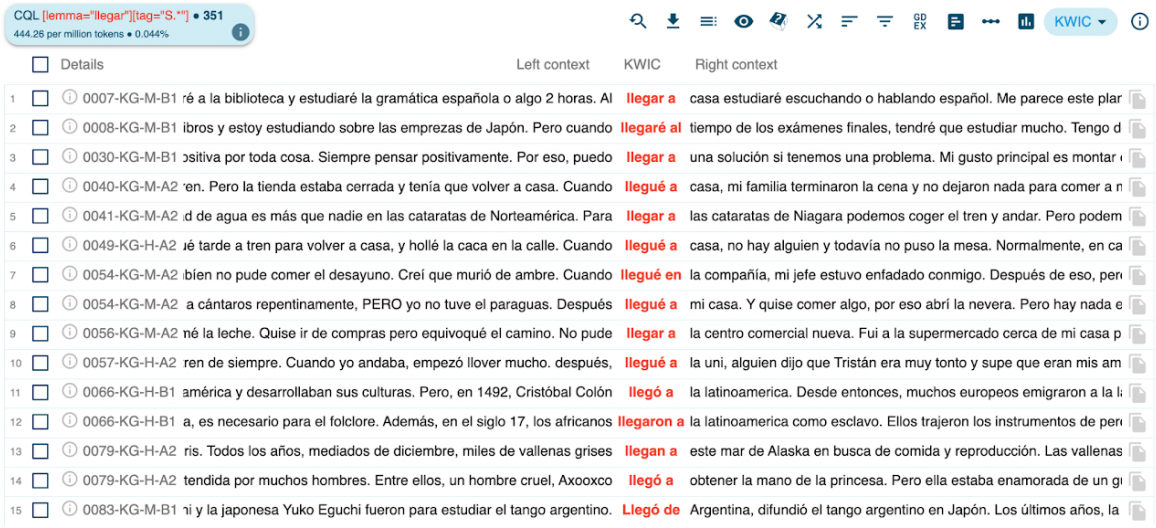
ISSN 2696-676X

Figura 6
Búsqueda avanzada con CQL: el lema llegar *seguido por una preposición*



Cada fragmento entre corchetes representa una palabra (un *token*). El punto indica "cualquier carácter" y el asterisco "el carácter anterior cero o más veces". Por lo tanto, tag="S.*" significa "una etiqueta que empieza por S", es decir, una preposición[12]. Con ello se obtienen 351 ejemplos, como muestra la Figura 7.

Figura 7
Concordancias del verbo (lema) llegar inmediatamente *seguido por una preposición (351 ocurrencias)*



Para resumir los resultados, simplemente hay que hacer clic en el botón *Frequency* y a continuación en *KWIC Lemmas*, con lo que se obtiene la lista de frecuencia de la Figura 8. Se pueden consultar directamente los ejemplos correspondientes a cada combinación mediante el menú local, en el icono de tres puntos (**...**) de la derecha. Ahí se observa que se usa *llegar en* como sinónimo de *llegar a*, tal vez por influencia del inglés, o por analogía con expresiones como *llegar en tren*.

[12] La lista de etiquetas puede consultarse en: https://sites.google.com/view/celen/documentación/etiquetas/

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

Figura 8
Lista de frecuencia del verbo llegar *más preposición (lemas)*

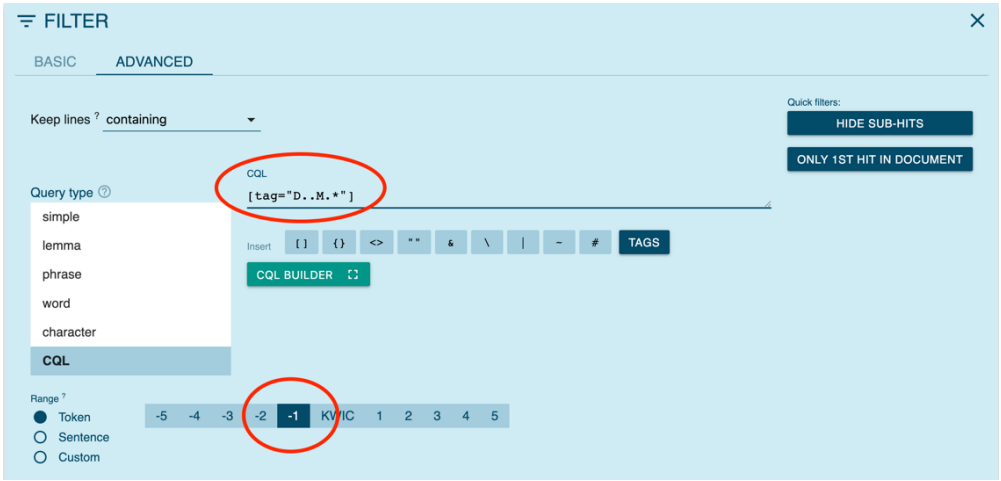
(9 items, 351 total frequency)

	Lemma	Frequency	Relative
1	llegar a	324	410.08
2	llegar en	10	12.66
3	llegar hasta	8	10.13
4	llegar de	3	3.80
5	llegar con	2	2.53
6	llegar durante	1	1.27
7	llegar desde	1	1.27
8	llegar sin	1	1.27
9	llegar tras	1	1.27

Otra de las áreas más problemáticas para los estudiantes japoneses es la **concordancia de género** y número, tanto en el sintagma nominal como en el verbal, ya que en su lengua no existe el género ni el número salvo en los pronombres. Para averiguar cuáles son los sustantivos femeninos más afectados por errores de concordancia de género, la siguiente expresión recuperará todos los sustantivos femeninos: `[tag="N.F.*"]`

A continuación, con el botón *Filter* (Figura 9) de la barra de herramientas, se pueden extraer sólo aquellas ocurrencias en las que el sustantivo vaya precedido de un determinante masculino (`D..M.*`), cuyo resultado muestra la Figura 10.

Figura 9
Filtro: se conservan solo las líneas de concordancia que contengan un determinante masculino (D..M.)* inmediatamente delante de la palabra buscada (-1 KWIC).



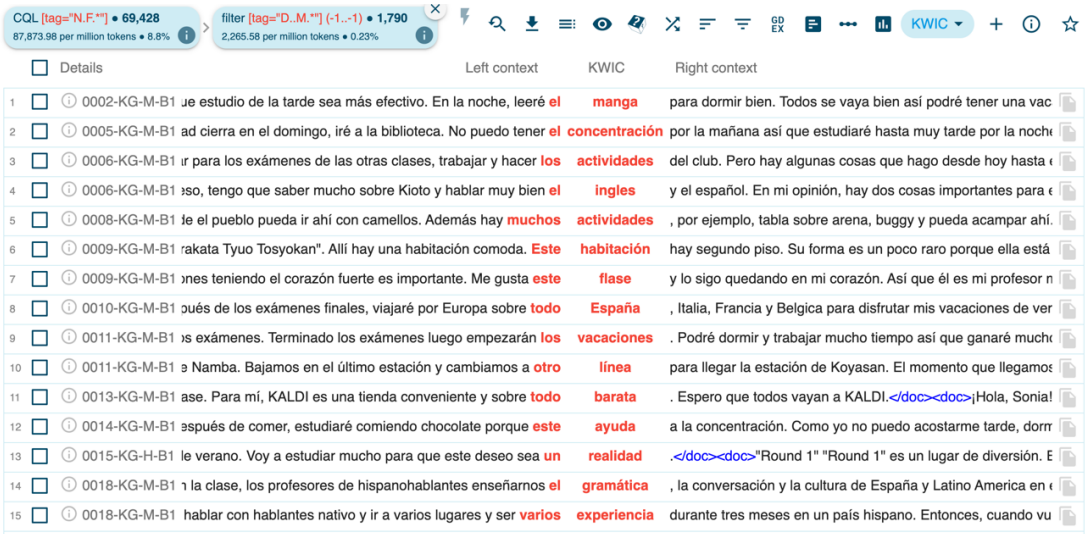
Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

Figura 10

Concordancias de un determinante masculino inmediatamente seguido por un sustantivo femenino (1.790 ocurrencias)



	Details	Left context	KWIC	Right context
1	0002-KG-M-B1	ue estudio de la tarde sea más efectivo. En la noche, leeré el	manga	para dormir bien. Todos se vaya bien así podré tener una vac
2	0005-KG-M-B1	ad cierra en el domingo, iré a la biblioteca. No puedo tener el	concentración	por la mañana así que estudiaré hasta muy tarde por la noch
3	0006-KG-M-B1	ır para los exámenes de las otras clases, trabajar y hacer los	actividades	del club. Pero hay algunas cosas que hago desde hoy hasta
4	0006-KG-M-B1	ıso, tengo que saber mucho sobre Kioto y hablar muy bien el	ingles	y el español. En mi opinión, hay dos cosas importantes para
5	0008-KG-M-B1	le el pueblo pueda ir ahí con camellos. Además hay muchos	actividades	, por ejemplo, tabla sobre arena, buggy y pueda acampar ahí.
6	0009-KG-M-B1	rakata Tyuo Tosyokan". Allí hay una habitación comoda. Este	habitación	hay segundo piso. Su forma es un poco raro porque ella está
7	0009-KG-M-B1	nes teniendo el corazón fuerte es importante. Me gusta este	flase	y lo sigo quedando en mi corazón. Así que él es mi profesor n
8	0010-KG-M-B1	ués de los exámenes finales, viajaré por Europa sobre todo	España	, Italia, Francia y Belgica para disfrutar mis vacaciones de ver
9	0011-KG-M-B1	s exámenes. Terminado los exámenes luego empezarán los	vacaciones	. Podré dormir y trabajar mucho tiempo así que ganaré much
10	0011-KG-M-B1	a Namba. Bajamos en el último estación y cambiamos a otro	línea	para llegar la estación de Koyasan. El momento que llegamos
11	0013-KG-M-B1	ase. Para mí, KALDI es una tienda conveniente y sobre todo	barata	. Espero que todos vayan a KALDI.</doc><doc>¡Hola, Sonia!
12	0014-KG-M-B1	spués de comer, estudiaré comiendo chocolate porque este	ayuda	a la concentración. Como yo no puedo acostarme tarde, dorm
13	0015-KG-H-B1	le verano. Voy a estudiar mucho para que este deseo sea un	realidad	.</doc><doc>"Round 1" "Round 1" es un lugar de diversión. E
14	0018-KG-M-B1	ı la clase, los profesores de hispanohablantes enseñarnos el	gramática	, la conversación y la cultura de España y Latino America en
15	0018-KG-M-B1	hablar con hablantes nativo y ir a varios lugares y ser varios	experiencia	durante tres meses en un país hispano. Entonces, cuando vu

Para ver la frecuencia de este fenómeno en cada nivel, hay que hacer clic en el botón *Frequency* de la barra de herramientas y a continuación en *Text types*. Ahí se observa que estos errores son más frecuentes en el nivel inicial, como es esperable (Figura 11).

Figura 11

Lista de frecuencia de "determinante masculino + sustantivo femenino" según el nivel de dominio del MCER. Frecuencia absoluta y frecuencia por millón.

(8 items, 2,869 total frequency)

	18. (A) Nivel (MCER)	Frequency	Relative in text type
1	A	1,184	2,770.01
2	A>A1	537	2,285.96
3	B	503	2,066.57
4	A>A2	379	2,737.71
5	B>B1	114	2,191.00
6	C	103	863.72
7	B>B2	44	2,377.09
8	C>C2	5	209.17

Para extraer la lista de frecuencias de los sustantivos femeninos, hay que hacer clic de nuevo en el botón *Frequency* y seleccionar la opción *KWIC lemmas*. Ahí veremos que los lemas más frecuentes son *costumbre, persona, gente, ciudad, cosa, carne, clase, foto, comida, vez, flor, razón, tienda, vacaciones,* etc. También hay algunos usos correctos (*área*) y errores de etiquetado (*dia*, sin acento, y *CD*). Esta información es de gran utilidad para profesores y alumnos, ya que los libros de texto suelen señalar sólo las excepciones a la regla general (sustantivos masculinos terminados en *a*, y femeninos terminados en *o*), pero no los errores más frecuentes. Desde el punto de vista de la investigación, sería interesante averiguar en qué circunstancias -con qué tipos de determinante y en qué tipos de sintagmas nominales- se producen más errores y por qué. Por ejemplo, un 10% de los errores anteriores se producen con el cuantificador *mucho*

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

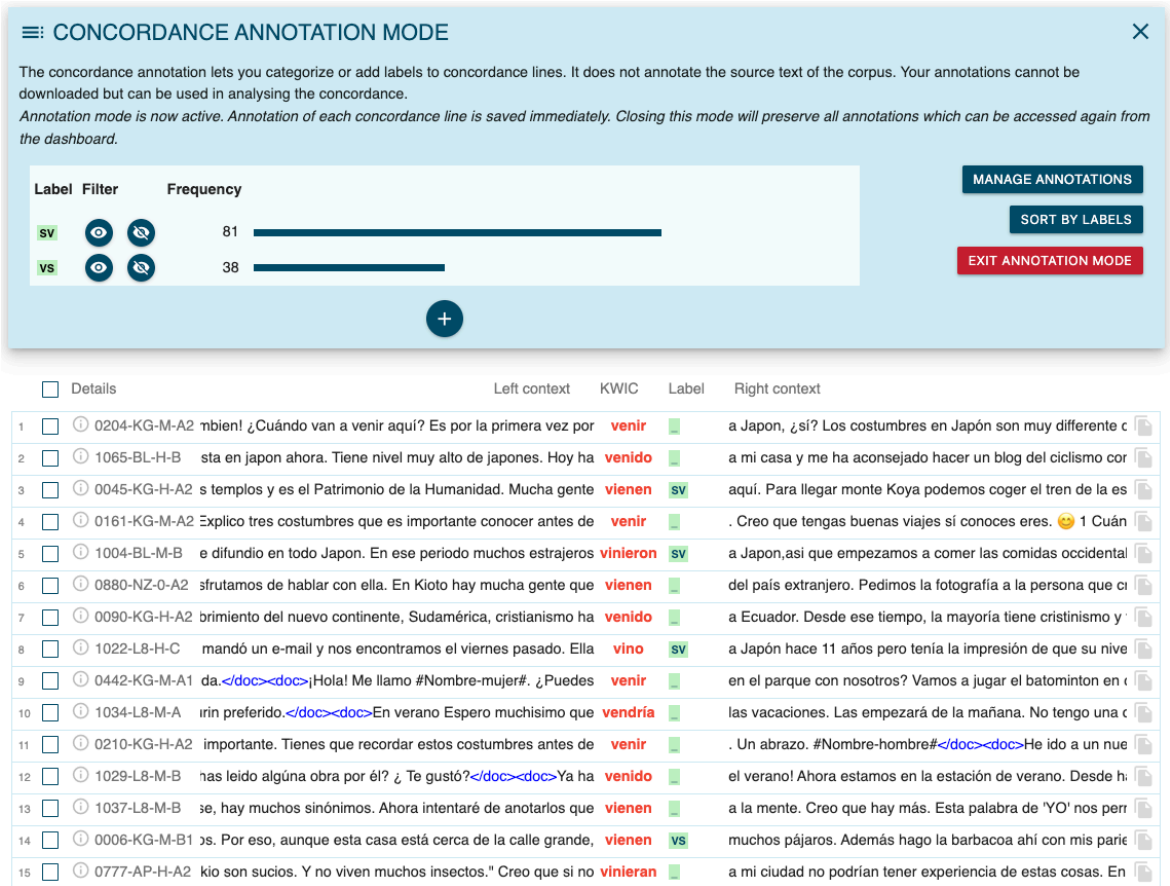
ISSN 2696-676X

(*mucho gente, muchos cosas, muchos personas…),* pero algunos de estos casos podrían deberse a una confusión entre el uso de esta palabra como adverbio o como determinante más que a una simple falta de concordancia.

Respecto al **orden de palabras**, el japonés es una lengua SOV, aunque sobre todo en la lengua hablada tiene bastante libertad. En español, los nativos suelen producir el orden SV con verbos inergativos (*hablar, trabajar, caminar…*) pero VS con verbos inacusativos (*existir, aparecer, llegar, venir…*). Para determinar si los aprendices japoneses son sensibles a esta distinción, se puede empezar por extraer ejemplos de un verbo inacusativo como *venir* (511 ocurrencias) y anotar las líneas de concordancia manualmente, con la herramienta *Concordance annotation mode*, con dos etiquetas: SV (sujeto antepuesto) y VS (sujeto pospuesto). Excluyendo los pronombres relativos antepuestos al verbo, encontramos 119 ejemplos de *venir* con un sujeto explícito, de los cuales 81 aparecen a la izquierda del verbo y 38 a la derecha (Figura 12). Por lo tanto, parece que en el corpus, los aprendices prefieren la anteposición del sujeto con el verbo *venir*, aunque por niveles, se observa un aumento gradual de VS del nivel A (8 de 41 casos, 20%) al B (22 de 64 casos, 34%) y el C (8 de 14 casos, 57%), donde VS se convierte en el orden predominante.

Figura 12

Anotación de las líneas de concordancia del verbo venir *con sujeto antepuesto (SV) o pospuesto (VS)*



Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

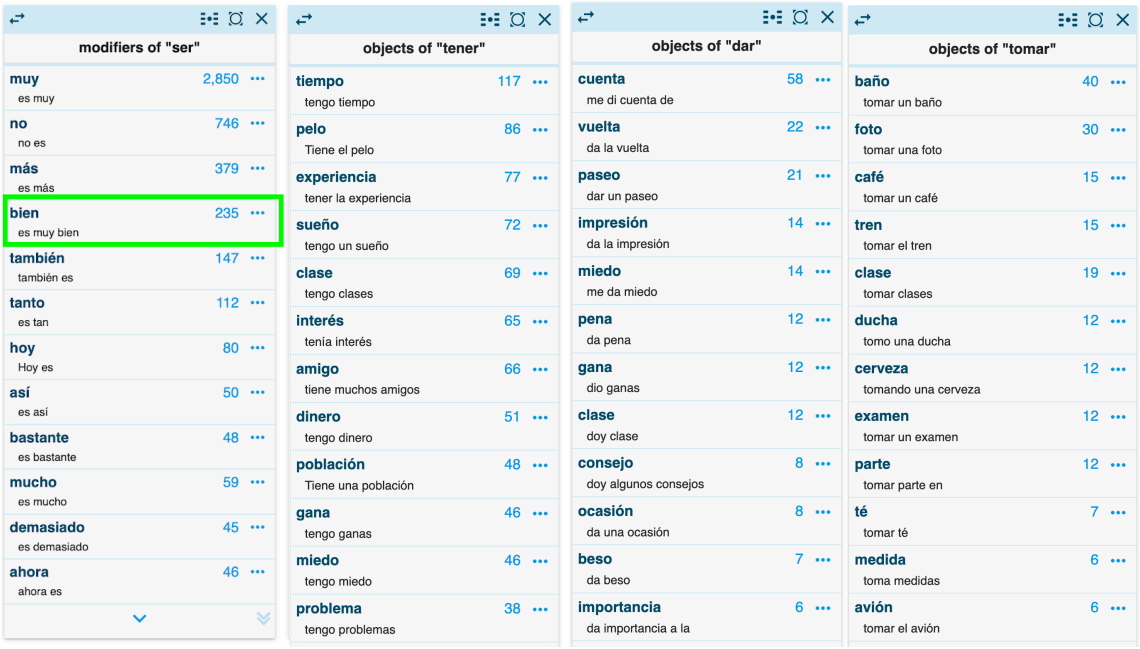
ISSN 2696-676X

3.3 Herramienta *Colocaciones*

En el proceso de enseñanza/aprendizaje de una lengua extranjera, es fundamental el dominio de las colocaciones (Lewis, 2000). Para el profesor, es importante saber qué colocaciones dominan y qué colocaciones necesitan aprender sus estudiantes, para tomar decisiones sobre los contenidos del curso. Sin embargo, apenas hay estudios sobre frecuencias ni descripciones de colocaciones léxicas por niveles de competencia (Blanco y Ferreira, 2021).

Con la herramienta *Word Sketch* se pueden extraer las colocaciones de cualquier palabra, teniendo en cuenta que son necesarios muchos ejemplos para obtener información fiable sobre su combinatoria y por lo tanto conviene limitar las búsquedas a los verbos, sustantivos o adjetivos más frecuentes, con cientos o miles de apariciones. En la Figura 13 se muestran algunas colocaciones de los verbos *ser* (20.430 apariciones), *tener* (6.074), *dar* (769) y *tomar* (483). Los resultados se agrupan según su función sintáctica, lo que permite ver rápidamente tanto los usos prototípicos (*tener tiempo, darse cuenta, tomar un baño*) como los usos no normativos pero muy frecuentes entre los aprendices como *ser bien*.

Figura 13
Colocaciones de los verbos ser, tener, dar *y* tomar *(lista parcial)*



modifiers of "ser"		objects of "tener"		objects of "dar"		objects of "tomar"	
muy (es muy)	2,850	tiempo (tengo tiempo)	117	cuenta (me di cuenta de)	58	baño (tomar un baño)	40
no (no es)	746	pelo (Tiene el pelo)	86	vuelta (da la vuelta)	22	foto (tomar una foto)	30
más (es más)	379	experiencia (tener la experiencia)	77	paseo (dar un paseo)	21	café (tomar un café)	15
bien (es muy bien)	235	sueño (tengo un sueño)	72	impresión (da la impresión)	14	tren (tomar el tren)	15
también (también es)	147	clase (tengo clases)	69	miedo (me da miedo)	14	clase (tomar clases)	19
tanto (es tan)	112	interés (tenía interés)	65	pena (da pena)	12	ducha (tomo una ducha)	12
hoy (Hoy es)	80	amigo (tiene muchos amigos)	66	gana (dio ganas)	12	cerveza (tomando una cerveza)	12
así (es así)	50	dinero (tengo dinero)	51	clase (doy clase)	12	examen (tomar un examen)	12
bastante (es bastante)	48	población (Tiene una población)	48	consejo (doy algunos consejos)	8	parte (tomar parte en)	12
mucho (es mucho)	59	gana (tengo ganas)	46	ocasión (da una ocasión)	8	té (tomar té)	7
demasiado (es demasiado)	45	miedo (tengo miedo)	46	beso (da beso)	7	medida (toma medidas)	6
ahora (ahora es)	46	problema (tengo problemas)	38	importancia (da importancia a la)	6	avión (tomar el avión)	6

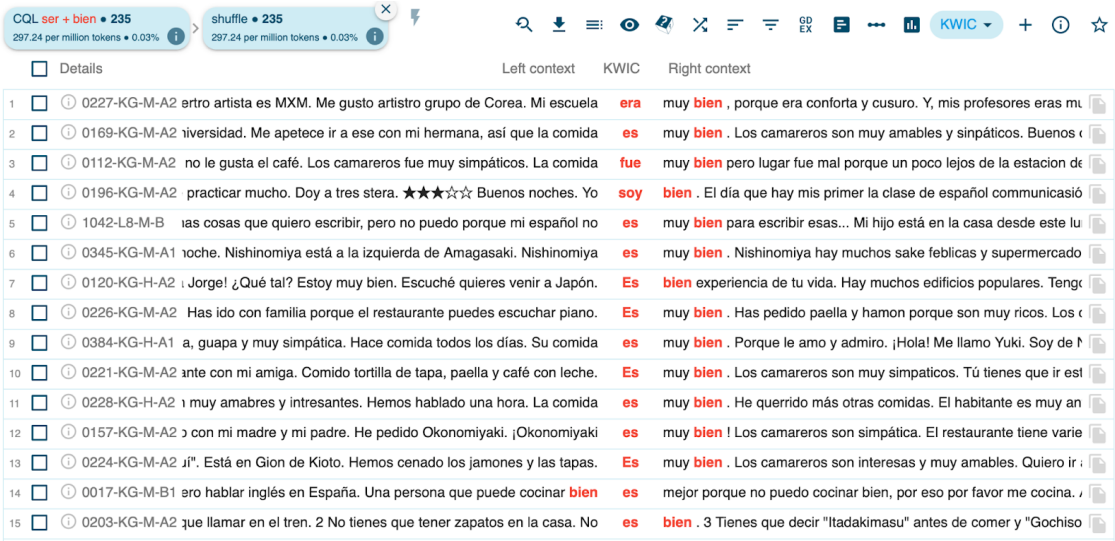
Nótese que al extraer las colocaciones no se tienen en cuenta simplemente las palabras inmediatamente a la derecha y a la izquierda del colocado sino que se realiza un análisis sintáctico superficial de la frase para admitir modificadores como determinantes o adverbios (Figura 14).

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

Figura 14
Concordancias de la colocación ser bien *(235 ocurrencias)*



La forma *ser bien* es especialmente común entre los aprendices japoneses, como expresión de valoración (de una comida, un lugar, una persona…). La confusión entre *ser* y *estar* se debe principalmente a la influencia de la lengua materna, ya que en japonés solo existe un verbo copulativo, mientras que en español hay dos, y suele estar relacionada con problemas de léxico: en español decimos *¡Qué bien!,* pero no *es bien* sino *está bien* o *es bueno*. De la misma manera, decimos *estoy triste,* pero no *estoy divertido* sino *lo estoy pasando bien* o *me estoy divirtiendo*.

3.4 Herramienta *Listas*

Con la herramienta *Listas* es posible obtener listas de frecuencia de palabras, etiquetas, lemas o cualquier combinación de atributos, según diversos criterios. Una aplicación inmediata de esta herramienta es la extracción del vocabulario que realmente usan los estudiantes. Como es sabido, los inventarios del *Plan Curricular* del Instituto Cervantes (2007) se basaron en la intuición de los autores y en la experiencia docente, pero no en datos de corpus. El inventario de nociones específicas incluye una lista de unidades léxicas clasificadas en veinte temas con varios subapartados (partes del cuerpo, relaciones familiares, alimentos…) pero no se trata de una lista exhaustiva sino un inventario abierto que, tras casi 20 años desde su publicación, necesita ser revisado y actualizado con datos basados en corpus.

En el caso de los aprendices japoneses, por ejemplo, podría averiguarse cuáles son los adjetivos más frecuentes en el nivel inicial. Para extraer esta lista de frecuencia simplemente hay que ir a la herramienta *Listas* y seleccionar el tipo de palabra y el nivel en el menú de búsqueda avanzada. Resulta obvio que el léxico empleado depende de los temas sobre los que tratan las tareas. Para atenuar este inconveniente, se pueden mostrar varios tipos de frecuencias (Figura 15).

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

Figura 15

Lista de frecuencia de los diez adjetivos (lemas) más frecuentes en el nivel A, de mayor a menor ARF

	Adjective	Frequency ?	Frequency Per Million ?	Relative DOCF ?	ARF ? ↓	
1	bueno	820	1,918.42	200.00 %	459.00	•••
2	grande	728	1,703.18	200.00 %	412.00	•••
3	importante	625	1,462.21	200.00 %	311.00	•••
4	primero	492	1,151.05	200.00 %	290.00	•••
5	famoso	479	1,120.64	200.00 %	265.00	•••
6	alto	496	1,160.41	200.00 %	260.00	•••
7	japonés	441	1,031.73	200.00 %	250.00	•••
8	español	509	1,190.82	200.00 %	246.00	•••
9	medio	510	1,193.16	200.00 %	223.00	•••
10	blanco	411	961.55	200.00 %	192.00	•••

Además de la frecuencia absoluta o la frecuencia por millón, puede ser especialmente útil la frecuencia ARF (*average reduced frequency*), en la columna de la derecha, que indica si la palabra se usa de forma homogénea en varias partes del corpus o, por el contrario, se concentra en alguna parte (en algún tipo de texto) en particular, en cuyo caso la cifra se alejará de la frecuencia absoluta[13]. Si se ordenan los resultados según ARF en lugar de la frecuencia absoluta, los diez adjetivos más frecuentes y de uso más general (más homogéneo) en el nivel A son, en este orden: *bueno, grande, importante, primero* (como en *primera vez*), *famoso, alto, japonés, español, medio* (como en *las siete y media*) y *blanco*.

3.5 Herramienta *N-gramas*

Las secuencias de categorías gramaticales (*POS-tags*) y su frecuencia han sido objeto de estudio en el campo de la adquisición de lenguas, como un medio para caracterizar el lenguaje escrito de los aprendices en distintos estadios de aprendizaje (Lim *et al.*, 2024). Con la herramienta *N-grams* se puede extraer, por ejemplo, una lista de las secuencias de cuatro categorías gramaticales junto a su frecuencia, por niveles. Para ello, hay que ir a la pestaña de búsqueda avanzada (Figura 16), especificar la longitud de la secuencia deseada (4) y el atributo (la etiqueta de categoría gramatical se llama *shorttag*). Para excluir los signos de puntuación (*F*), hay que escribir la letra efe en la casilla correspondiente. Por último, en el menú *Text types* se puede seleccionar el nivel de dominio.

[13] Para más detalles sobre la manera de calcular ARF, véase https://www.sketchengine.eu/documentation/average-reduced-frequency/

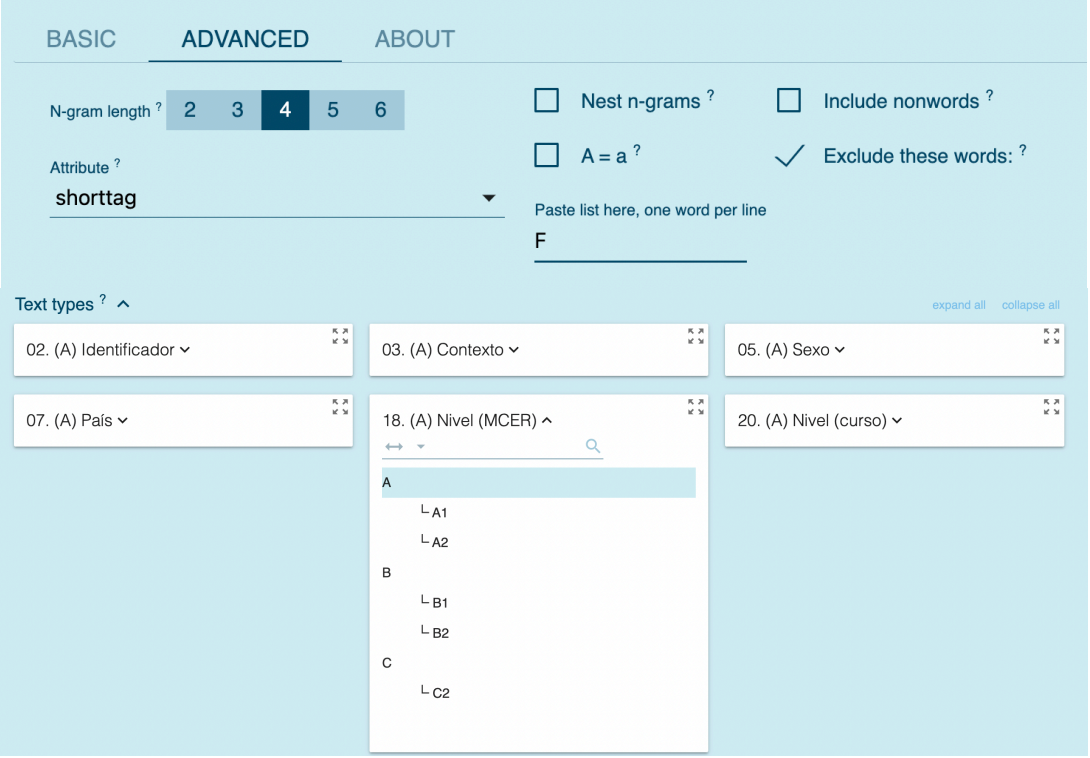
Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

Figura 16

*Condiciones de búsqueda de n-gramas formados por 4 categorías gramaticales (*shorttag*), excluyendo signos de puntuación (F), en textos de nivel A*



En la Figura 17 vemos que, entre los aprendices japoneses, la tendencia más destacable es el aumento de frecuencia de algunos n-gramas del nivel A a los niveles B y C, lo que podría atribuirse al aumento en el grado de complejidad del sintagma nominal (NSDN, DNSD, VDNS) y el sintagma verbal (VVDN). Mientras que en el nivel A la secuencia nombre-preposición-determinante-nombre (NSDN) suele darse en estructuras como *(tener) clase por la mañana*, en los niveles B y C se corresponde más comúnmente con un sintagma nominal complejo (*menú de la cena, gracias por su colaboración*). Lo mismo ocurre con la secuencia determinante-nombre-preposición-determinante (DNSD), que en el nivel A es a menudo del tipo (*voy a*) *la universidad a las (nueve),* mientras que en los niveles B y C de nuevo se corresponde con un sintagma nominal complejo (*el menú de la, la mayoría de los*). Del mismo modo, también aumenta la frecuencia de verbo-determinante-nombre-preposición (VDNS) a medida que aumenta el nivel de dominio (*lleva una camisa de, es el equipo de, espero sus comentarios al*). La secuencia verbo-verbo-determinante-nombre (VVDN) también aumenta su frecuencia desde el nivel A (*he ido un restaurante*) a los niveles B y C (*estoy buscando una persona, estaba leyendo un artículo*), lo que indica un mayor uso de perífrasis verbales o tiempos compuestos en los niveles intermedio y alto.

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

Figura 17

Lista de frecuencias parcial de n-gramas (4 elementos, categorías gramaticales), en los niveles A (izquierda), B (centro) y C (derecha) del MCER. D = determinante, N = sustantivo, S = preposición, V = verbo, P = pronombre, A = adjetivo

	N-gram	Frequency	Frequency per million
1	DNSN	6,793	15,892.44
2	VSDN	4,684	10,958.37
3	NSDN	4,523	10,581.70
4	VDNS	3,921	9,173.30
5	SDNS	3,515	8,223.45
6	DNSD	2,817	6,590.46
7	PVDN	2,467	5,771.62
8	SNSN	2,319	5,425.37
9	NSNS	1,933	4,522.31
10	VDNA	1,882	4,403.00

	N-gram	Frequency	Frequency per million
1	DNSN	3,461	14,219.51
2	NSDN	3,189	13,102.00
3	VDNS	2,896	11,898.21
4	VSDN	2,761	11,343.56
5	DNSD	2,043	8,393.66
6	SDNS	1,978	8,126.61
7	VVDN	1,422	5,842.28
8	PVDN	1,398	5,743.68
9	DNPV	1,263	5,189.03
10	VDNA	1,179	4,843.92

	N-gram	Frequency	Frequency per million
1	NSDN	1,672	14,020.73
2	DNSN	1,653	13,861.40
3	VSDN	1,372	11,505.05
4	VDNS	1,325	11,110.92
5	SDNS	1,148	9,626.67
6	DNSD	1,093	9,165.46
7	PVDN	661	5,542.88
8	NSNS	653	5,475.80
9	VVDN	627	5,257.77
10	SNSN	604	5,064.90

Conclusiones y líneas futuras

El corpus CELEN aporta una muestra amplia y diversa de la expresión escrita de los aprendices japoneses y está a disposición de la comunidad investigadora mediante una aplicación de consulta potente e intuitiva. Con ello, esperamos que contribuya a la difusión de la lingüística de corpus en el campo de la docencia e investigación en ELE.

En la enseñanza, una de las aplicaciones más inmediatas de este recurso es la especificación de los elementos de la lengua que realmente producen los aprendices en cada nivel. Para ello suele tomarse como referencia el Marco Común Europeo de Referencia (MCER) y las especificaciones del *Plan Curricular* del Instituto Cervantes (2007) o, en el contexto japonés, el *Modelo de contenidos* del grupo GIDE (2015). Todos ellos fueron elaborados mediante procedimientos intuitivos (Llorián, 2017) y por lo tanto sería beneficioso averiguar qué contenidos gramaticales o léxicos se usan realmente en cada nivel y con qué valores o grado de corrección, como se ha hecho con el inglés a un nivel mucho mayor en el proyecto *Learner profile* (https://www.englishprofile.org/). Los resultados de un estudio de este tipo servirán para tomar decisiones en la planificación curricular y desarrollar materiales adaptados a las necesidades de los aprendices japoneses.

En la investigación, los corpus de aprendices son de gran ayuda en varios campos relacionados con el aprendizaje de lenguas: en el análisis de errores (Saito, 2005; Fernández, 1997, *inter alia*), para descubrir las causas de los errores más comunes y las estrategias que los alumnos usan durante su aprendizaje; en la lingüística contrastiva, cuando se comparan dos o más lenguas –el español y el japonés–, para determinar las diferencias y las semejanzas existentes entre ellas (Romero Díaz, 2011; Fukushima, 2014; Sanz *et al.*, 2015; Civit i Contra, 2016; Takagaki, 2018, *inter alia*); o en la lingüística cognitiva, para observar por ejemplo cómo se desarrollan la metáfora y la metonimia (Suárez-Campos y Hijazo-Gascón, 2019).

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

A través de ejemplos prácticos, se ha mostrado cómo se puede explotar el corpus para la investigación en ELE. Mediante la herramienta *Concordancias* es posible recuperar ejemplos de acuerdo con varios criterios, como hemos ejemplificado para los verbos *caer(se), llegar* o *venir*, así como anotar las líneas de concordancia. Gracias a la anotación subyacente, es posible realizar búsquedas complejas sobre la etiqueta de las palabras, y recuperar por ejemplo errores en la concordancia de género entre determinante y sustantivo. El programa de consulta dispone también de otras herramientas para extraer listas de palabras, n-gramas o palabras semejantes, entre otros. En la página del proyecto, https://sites.google.com/view/celen, puede consultarse la guía de uso detallada, donde se explican otras herramientas para extraer palabras clave (*Keywords*), palabras semejantes (*Thesaurus*), diferencias entre dos palabras (*Word Sketch Difference*), estadísticas sobre todo el corpus, etc., que por motivos de espacio no se comentan en este artículo. En la misma página también pueden descargarse íntegramente algunas partes del corpus bajo una licencia CC BY-NC 4.0.

Como líneas de mejora futuras, además de ampliar el corpus de forma periódica, está previsto revisar de manera semiautomática la anotación morfosintáctica de los lemas más frecuentes, asignar de forma manual algunos metadatos pendientes a los textos procedentes de Internet e incorporar datos de producción oral como los de García Ruiz-Castillo (2022). A más largo plazo, sería conveniente complementar el corpus de aprendices con datos comparables de nativos, para poder hacer comparaciones aprendiz-nativo. Se trata de un corpus abierto y esperamos que otros profesores e investigadores alberguen sus textos en él para ofrecer a la comunidad científica una amplia muestra de aprendices japoneses de español.

Agradecimientos

Esta investigación ha sido financiada por *kakenhi* (17H07270 y 23K00698) *Grant-in-Aid for Scientific Research* de la Japan Society for the Promotion of Science. El desarrollo del corpus CELEN ha sido posible gracias a la colaboración inestimable de numerosos profesores y alumnos. Damos las gracias especialmente al profesor Nobuyuki Tukahara, de la Universidad de Kioto, por ceder los datos correspondientes a dicha universidad; a la profesora Atsuko Wasa y el profesor Muneaki Tsujii de la Universidad Kansai Gaidai, por su valiosa colaboración; y al profesor Yoshihito Kamakura de la Universidad de Aichi, por ceder los datos del corpus JALCOS.

Referencias

Alexopoulou, Theodora; Meurers, Detmar; Murakami, Akira (2022). Big data in SLA: advances in methodology and analysis. En Nicole Ziegler; Marta González-Lloret (Eds.), *The Routledge handbook of second language acquisition and technology,* pp. 92–106. Routledge. https://doi.org/10.4324/9781351117586-9

Alsonso-Ramos, Margarita (2016). *Spanish Learner Corpus Research. Current Trends and Future Perspectives*. John Benjamins. https://doi.org/10.1075/scl.78

Atkins, Sue; Clear, Jeremy; Oster, Nicholas (1992). Corpus Design Criteria. *Literary and Linguistic Computing* 7 (1), 1–16. https://doi.org/10.1093/llc/7.1.1

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

© La autora | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

Badillo Matos, Ángel (2021). *Lengua y cultura en español en el Japón de la era Reiwa*. Instituto Cervantes y Fundación Real Instituto Elcano de Estudios Internacionales y Estratégicos. https://media.realinstitutoelcano.org/wp-content/uploads/2021/01/badillo-lengua-y-cultura-en-espanol-en-japon-era-reiwa-1.pdf

Berdicevskis, Aleksandrs (2020). Foreigner-directed speech is simpler than native-directed: Evidence from social media. En *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pp. 163–172. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.nlpcss-1.18

Bailini, Sonia; Frigerio, Aldo (2019). CORESPI y CORITE: criterios de construcción e implementación en línea de dos corpus de interlengua paralelos. *E-AESLA 5*, 303–312. https://cvc.cervantes.es/lengua/eaesla/pdf/05/29.pdf

Blanco, Lorena; Ferreira, Anita (2021). La importancia de las colocaciones léxicas verbonominales en aprendientes de ELE. RLA. *Revista de lingüística teórica y aplicada*, *59*(1), 91–112. https://dx.doi.org/10.29393/rla59-4icla20004

Brooke, Julian; Hirst, Graeme (2013). Native language detection with 'cheap' learner corpora. En Sylviane Granger, Gaëtanelle Gilquin & Fanny Meunier (Eds.), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Corpora and Language in Use – Proceedings 1,* pp. 37–47. Presses Universitaires de Louvain. http://ftp.cs.toronto.edu/pub/gh/Brooke+Hirst-LCRbook-2013.pdf

Buyse Kris; González Melón, Eva (2013). El corpus de aprendices Aprescrilov y su utilidad para la didáctica de ELE en la Bélgica multilingüe. En *Plurilingüismo y enseñanza de ELE en contextos multiculturales: XXIII Congreso Internacional ASELE*, 247–252. https://cvc.cervantes.es/ensenanza/biblioteca_ele/asele/pdf/23/23_0025.pdf

Callies, Marcus; Paquot, Magali (2015). Learner Corpus Research: An interdisciplinary field on the move. *International Journal of Learner Corpus Research 1*(1), 1–6. https://doi.org/10.1075/ijlcr.1.1.00edi

Campillos Llanos, Leonardo (2014). A Spanish learner oral corpus for computer aided error analysis. *Corpora* 9(2), 207–238. https://doi.org/10.3366/cor.2014.0058

Cassany, Daniel (2023). Aprender ELE (y otras L2) en contextos informales. *Cuadernos CANELA 33*, 5–23. https://doi.org/10.2107/canela.33.0_5

Cestero Mancera, Ana María; Penadés Martínez, Inmaculada; Blanco Canales, Ana; Camargo Fernández, Laura; Francisco Simón Granda, José (2001). Corpus para el análisis de errores de aprendices de E/LE (CORANE). En *Actas del XII Congreso Internacional de ASELE*, pp. 527–534. https://dialnet.unirioja.es/descarga/articulo/2553470.pdf

Civit i Contra, Roger (2016). Teoría de eventos y las expresiones no eventivas del español. *Cuadernos CANELA 27*, 110–126. https://cuadernoscanela.org/index.php/cuadernos/article/view/62

Corder, Stephen Pit (1981). *Error Analysis and Interlanguage*. Oxford University Press.

Cruz Piñol, Mar (2012). *Lingüística de corpus y enseñanza del español como 2/L.* Arco/Libros.

Davies, Mark; Biber, Douglas; Jones, James K.; Tracy-Ventura, Nicole (2006). Spoken and Written Register Variation in Spanish: A Multi-dimensional Analysis. *Corpora 1,* 1–37. https://www.mark-davies.org/articles/davies_30.pdf

Davies, Mark (2013). Establishing Corpora from Existing Data Sources. En Christine Mallinson; Becky Childs; Gerard Van Herk (Eds.), *Data Collection in Sociolinguistics: Methods and Applications*, pp. 210–212. Routledge. https://www.english-corpora.org/davies/articles/davies_53.pdf

Elvira-García, Wendy (2021). *Uso de corpus en clase de ELE: la lengua real como modelo*. Difusión.

Fernández, Sonsoles (1997). *Interlengua y Análisis de Errores en el aprendizaje del español como lengua extranjera*. Edelsa.

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

Fukushima, Noritaka (2014). *El español y el japonés*. Monograph series in Foreign Studies, 53. Kobe City University of Foreign Studies. http://id.nii.ac.jp/1085/00001678/

García Ruiz-Castillo, Carlos (2022). Creación de un corpus oral para el estudio de la conversación en español de aprendientes japoneses de ELE. *Hiroshima Studies in Language and Language Education 25*, 155–170. https://doi.org/10.15027/51967

GIDE (Grupo de Investigación de la Didáctica del Español) (2015). *Un modelo de contenidos para un modelo de actuación. Enseñar español como segunda lengua extranjera en Japón*. https://conchamorenogarcia.es/2016/01/30/ensena-espanol-en-japon-modelo-de-contenidos-de-gide/

Gilquin, Gaëtanelle (2015). From design to collection of learner corpora. En Sylviane Granger; Gaëtanelle Gilquin; Fanny Meunier (Eds.), *The Cambridge Handbook of Learner Research*, 9–34. Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.002

Hunston, Susan (2008). Collection Strategies and Design Decisions. En Anke Lüdeling; Merja Kytö (Eds.), *Corpus Linguistics: an International Handbook*, pp. 154–168,Walter de Gruyter.

Instituto Cervantes (2007). *Plan Curricular del Instituto Cervantes. Niveles de referencia para el español*. Biblioteca Nueva. https://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/indice.htm

Instituto Cervantes (2020). *El español en el mundo 2020*. Instituto Cervantes. https://cvc.cervantes.es/lengua/anuario/anuario_20/

Kilgarriff, Adam; Grefenstette, Gregory (2003). Introduction to the special issue on the Web as corpus. *Computational Linguistics 29*(3), 333–347. https://doi.org/10.1162/089120103322711569

Leech, Geoffrey (2007). New Resources or Just Better Old Ones? The Holy Grail of Representativeness. En Marianne Hundt; Nadja Nesselhauf; Carolin Biewer (Eds.), *Corpus Linguistics and the Web*, pp. 133–149. Rodopi. https://doi.org/10.1163/9789401203791_009

Lewis, Michael (1993). *The lexical approach*. LTP.

Lewis, Michael (2000). *Teaching collocation. Further developments in the lexical Approach*. LTP.

Lim, Joyce; Mark, Geraldine; Pérez-Paredes, Pascual; O'Keeffe, Anne (2024, en prensa). Exploring Part of Speech (POS)-tag sequences in a large-scale learner corpus of L2 English: A developmental perspective. *Corpora*, *19*(1). http://www.perezparedes.es/exploring-part-of-speech-pos-tag-sequences-in-a-large-scale-learner-corpus-of-l2-english-a-developmental-perspective/

Lozano, Cristóbal (2022). CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research*, 38(4), 965–983. https://doi.org/10.1177/02676583211050522

Llorián González, Susana (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*. https://marcoele.com/descargas/25/llorian_revision-nre.pdf

Lu, Hui Chuan (2010). An annotated Taiwanese learners' corpus of Spanish, CATE. *Corpus Linguistics and Linguistic Theory*, *6*(2), 297-300. https://doi.org/10.1515/CLLT.2010.011

McEnery, Tony; Hardie, Andrew (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press. http://doi.org/10.1017/CBO9780511981395

Mizumoto Tomoya; Komachi, Mamoru; Nagata, Masaaki; Matsumoto, Yuji (2011). Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners. En *Proceedings of the 5th*

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

International Joint Conference on Natural Language Processing (IJCNLP), pp. 147–155. https://aclanthology.org/I11-1017.pdf

Moreno, Concha (2022). Enseñar ELE en Japón. En María Méndez Santos; Mar Galindo Merino (Eds.), *Atlas de ELE. Geolingüística de la enseñanza del español en el mundo. Volumen II. Asia Oriental*, pp. 249–275. En Clave ELE. https://www.todoele.net/sites/default/files/atlas/10-japon_0.pdf

Nation, Ian Stephen Paul (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review/ La Revue canadienne des langues vivantes 63*(1), 59–81. https://doi.org/10.1353/cml.2006.0049

Padró, Lluís; Stanilovsky, Evgeny (2012). FreeLing 3.0: Towards Wider Multilinguality. En *Proceedings of the Language Resources and Evaluation Conference (LREC 2012).* ELRA Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2012/pdf/430_Paper.pdf

Rojo, Guillermo; Palacios, Ignacio; Sampedro, María; Marsily, Aurélie (2023). Los corpus de aprendices de español LE/L2: panorama actual y perspectivas futuras. *Journal of Spanish Language Teaching, 9*(2), 174–189. https://doi.org/10.1080/23247797.2022.2157085

Palacios Martínez, Ignacio; Barcala Rodríguez, Francisco Mario; Rojo, Guillermo (2019). El "Corpus de Aprendices del Español" (CAES) y sus aplicaciones para la enseñanza/aprendizaje del español como lengua extranjera. En Marta Blanco, Hella Olbertz y Victoria Vázquez Rozas (Eds.). *Corpus y construcciones. Perspectivas hispánicas. Anejo 79 de Verba*, pp. 273–303. https://dx.doi.org/10.15304/9788417595876

Romero Díaz, Juan (2011). Falta de correspondencia entre las categorías léxicas del español y el japonés y su influencia en la clase de ELE. En Ana Sánchez Urquijo (Ed.), *Competencias y estrategias docentes en el contexto de Asia Pacífico*, pp. 118–130. Instituto Cervantes de Manila y Embajada de España en Filipinas. https://cvc.cervantes.es/ensenanza/biblioteca_ele/publicaciones_centros/pdf/manila_2011/09_investigaciones_01.pdf

Saito, Akemi (2005). *Análisis de errores en la expresión escrita de los estudiantes japoneses*. Memoria de máster. Universidad de Salamanca. https://www.educacionyfp.gob.es/dam/jcr:bfd680f6-6dbc-41b7-8299-42cbed9c0c4a/2005-bv-03-13saito-pdf.pdf

Sanz, Montserrat; Escandón, Arturo; Romero Díaz, Juan; Ramírez Gómez, Danya; Civit i Contra, Roger (2015). Enseñar español en Japón. Algunos aspectos de la enseñanza a japoneses. *Annals of Foreign Studies 89*. Kobe City University of Foreign Studies http://id.nii.ac.jp/1085/00001825/

Schäfer, Roland; Bildhauer, Felix (2013). *Web Corpus Construction*. Springer Cham. https://doi.org/10.1007/978-3-031-02152-7

Sinclair, John (2005). Corpus and text – Basic principles. En Martin Wynne (Ed.), *Developing linguistic corpora: A guide to good practice*, pp. 1–16. Oxbow Books. https://users.ox.ac.uk/~martinw/dlc/chapter1.htm

Suárez-Campos, Laura; Hijazo-Gascón, Alberto (2019). La metáfora conceptual y su aplicación a la enseñanza del español LE/L2. En Iraide Ibarretxe-Antuñano; Teresa Cadierno; Alejandro Castañeda Castro (Eds.), *Lingüística cognitiva y español LE/L2,* pp. 235–252, Routledge.

Takagaki, Toshihiro (ed.) (2018) *Exploraciones de la lingüística contrastiva español-japonés*. Ediciones de la Universidad Autónoma de Madrid.

Tracy-Ventura, Nicole; Myles, Florence (2015). The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research 1* (1), 58–95. https://doi.org/10.1075/ijlcr.1.1.03tra

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023

ISSN 2696-676X

Yamada, Aaron; Davidson, Sam; Fernández-Mira, Paloma; Carand, Agustina; Sagae, Kenji; Sánchez-Gutiérrez, Claudia (2020), COWS-L2H A corpus of Spanish learner writing. *Research in Corpus Linguistics (RiCL) 8* (1), 17–32. https://doi.org/10.32714/ricl.08.01.02

Valverde, Pilar (2011). An evaluation of part of speech tagging on written second language Spanish. En Alexander Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing. CICLing 2011. Lecture Notes in Computer Science*, vol. 6609, pp. 214–226, Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-19400-9_17

Valverde, Pilar (2018). Un corpus de blogs de aprendices japoneses de español. En María Bargalló Escrivá; Esther Forgas Berdet; Antoni Nomdedeu Rull (Eds.), *Léxico y cultura en LE/L2: corpus y diccionarios. Actas del XXVIII Congreso Internacional de ASELE*, pp. 845–857. Asociación para la Enseñanza del Español como Lengua Extranjera. https://doi.org/10.13140/RG.2.2.35687.14246

Valverde, Pilar (2020). Diseño y creación de un corpus de aprendices de ELE en Japón (CELEN). *E-AESLA Revista digital de lingüística aplicada 6*, 223–240, Centro Virtual Cervantes. https://cvc.cervantes.es/lengua/eaesla/pdf/06/16.pdf

Valverde, Pilar (2023). El corpus de aprendices japoneses CELEN y su aplicación a la docencia y la investigación en ELE. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *3*, Artículo e-a_hr01, 1-31. http://doi.org/10.1344/teisel.v3.42898

 | Recepción: 21/05/2023 | Revisión: 16/06/2023| Aceptación: 18/10/2023| Publicación: 26/10/2023