Giovanni Parodi
Pontificia Universidad Católica de Valparaíso
gparodi@ucv.cl

Guillermo Rojo e Ignacio Palacios, dirs. ***Corpus de aprendices de español* (CAES)**. Instituto Cervantes. (Versión: 1.0 - octubre 2014). http://galvan.usc.es/caes

A mediados de la década de los 80, Johns (1986) propuso que una de las tareas de los aprendientes de una lengua extranjera debería ser "descubrir" la lengua, en lo que se ha conocido como "aprendizaje guiado por datos" (*data-driven learning,* DDL). En el núcleo de este enfoque yace el supuesto de que el uso del computador no sustituiría al profesor o tutor, sino que la máquina actuaría como un informante capaz de apoyar el aprendizaje con materiales lingüísticos auténticos. Desde esta idea original, mucho se ha avanzado en lo que ha venido a llamarse "corpus de o para aprendientes o aprendices" (del inglés *learner corpora*, Aijmer 2009, Granger, Gilquin y Meunier 2013). Todo lo cual, en palabras de Flowerdew (2015, 15), "has inspired a plethora of data-driven learning (DDL) initiatives across the globe". Incluso, el interés creciente en los corpus para aprendientes ha llegado a formalizarse en la constitución de una asociación internacional que aglutina a los investigadores en esta temática, la cual brinda aún mayor visibilidad al trabajo científico en este ámbito: *Learner Corpus Association*, LCA (http://www.learnercorpusassociation.org).

El impacto del innovador enfoque DDL, basado en técnicas clásicas desarrolladas y empleadas por investigadores de lingüística de corpus (Sinclair 1991, 1994, Baker 2009, Parodi 2010a, McEnery y Hardie 2011, Timmins 2015), ha conllevado un progresivo desarrollo de herramientas computacionales para el apoyo a la investigación así como de soporte para los procesos de enseñanza/aprendizaje no solo para lenguas extranjeras sino también para lenguas maternas. En este marco se inserta el presente análisis del *Corpus de aprendices de español* (CAES) (http://galvan.usc.es/caes), elaborado por el equipo liderado por Guillermo Rojo e Ignacio Palacios de la Universidad de Santiago de Compostela y promovido y financiado por el Instituto Cervantes.

El sitio web del CAES se compone de dos apartados fundamentales: una página inicial en la que se ofrece una cantidad importante de información contextual y la herramienta de búsqueda propiamente tal, denominada "aplicación de consulta".

En su página inicial, el CAES se define como "un conjunto de textos escritos producidos por estudiantes de español con diferentes grados de dominio lingüístico (niveles A1 a C1 del Marco común europeo de referencia, aplicado al español en el Plan curricular del Instituto Cervantes. Niveles de referencia para el español) y procedentes de seis L1: árabe, chino mandarín, francés, inglés, portugués y ruso". En esta definición quedan de manifiesto, sin lugar a dudas, algunas de las fortalezas y mayores riquezas de esta propuesta: la disponibilidad en línea de producción textual escrita de un número significativo de sujetos con diversos niveles de dominio lingüístico y proveniente de diversas lenguas maternas. En estos dos rasgos reside un aspecto muy innovador de los esfuerzos conjuntos de investigadores de la Universidad de Santiago de Compostela y del Instituto Cervantes, ciertamente una dupla robusta para alcanzar el objetivo propuesto.

En esta misma primera página del CAES se explicita que esta iniciativa constituye una herramienta que "permite a los profesionales del campo de ELE (profesores, investigadores, evaluadores, autores de materiales didácticos, responsables y equipos de centros e instituciones lingüísticas, etc.) llevar a cabo investigaciones aplicadas sobre la base de datos sólidos y objetivos, ya que puede proporcionar información sobre dificultades de aprendizaje, errores más comunes, vocabulario más o menos empleado, etc. que se podrá aplicar con facilidad en las aulas o integrar en los textos". En este sentido, este desarrollo tecnológico de gran envergadura deja

claro que, en su estado actual, el CAES no sería propiamente una herramienta para uso de estudiantes de español, ya fuera como lengua materna o extranjera. Por el contrario, se define como un apoyo a la investigación de especialistas profesionales. Así, el CAES no busca ser un "corpus *para* aprendientes" ni un "corpus *con* aprendientes", sino por el contrario —y como bien lo declaran— es un "corpus *de* aprendientes y *para* profesionales expertos". Valga esta aclaración pues nos resulta muy relevante redundar en la audiencia específica, ya que en el ámbito aún se detecta cierta variación terminológica a partir de la ambigüedad presente en la expresión original del inglés (*learner corpora*), en donde las preposiciones aclaratorias no se emplean (Granger, Hung y Petch-Tyson 2002, Gavioli 2005, Facchinetti 2007, Meunier, De Cock, Gilquin y Paquot 2011).

En su versión 1.0 (octubre de 2014) el CAES, tal como se declara en el sitio web, incluye cerca de 575.000 elementos lingüísticos, con una distribución que abarca un conjunto muy variado de niveles de dominio lingüístico y con atención a una diversidad importante de lenguas maternas (seis en total, como ya se señaló). Para su construcción, el equipo a cargo compiló un grupo de textos provenientes de un conjunto de centros del Instituto Cervantes en distintos lugares del mundo y de diversas universidades de un gran número de países. En esta primera fase del proyecto, se consideró un corpus en un período de tiempo que abarca desde el mes de octubre de 2011 hasta septiembre de 2013, el cual seguramente será de tipo incremental. Se hace relevante comentar que el corpus original de pruebas escritas recolectadas fue mayor al que hoy se presenta disponible. El equipo responsable ejecutó un proceso de revisión y filtrado y decidió, en base a diversas razones, eliminar algunos textos. En este escenario, la versión del CAES que comentamos se constituye de un corpus de 3.878 tareas integradas en 1.423 pruebas, producidas por 1.423 estudiantes, quienes escribieron dos o tres textos cada uno (según los niveles aprobados por los estudiantes escritores). Se hace especial mención al hecho de que no se aplicó ningún sistema de corrección o ajuste a los textos que integran la muestra, sino solo un proceso de desambiguación manual que permitiera el posterior etiquetaje morfosintáctico automático. Adicionalmente, para mayores detalles acerca del tipo de tarea que dio lugar a las producciones escritas de los estudiantes de español desde entornos diversos y su encuadre metodológico y contextual, el sitio web aporta información valiosa que permite construir una visión de conjunto de la procedencia de los textos y de las fortalezas y limitaciones que ellos encierran. Al respecto, se recomienda revisar el documento "CAES: Tipos de tareas escritas", disponible en el enlace "textos".

En cuanto a la información contextual que se puede explorar desde su página de inicio, recomendamos especialmente el apartado de "Distribución de las muestras" (disponible en el enlace "distribución"), en el cual se registran datos pormenorizados de los sujetos escritores de los textos que conforman el corpus. También sugerimos consultar el archivo descargable en formato PDF con el etiquetario desarrollado y empleado en el proyecto denominado "Etiquetario CAES. Versión 1.0" (disponible en el enlace "sistema de categorías y subcategorías"). Ambos apartados, entre otros datos disponibles, proporcionan una amplia gama de detalles que permite diseñar investigaciones y explorar aplicaciones potenciales. Asimismo, en el hipervínculo "Documentación complementaria", se registra el acceso a cinco documentos o secciones: 1) Formas y lemas del CAES, 2) Estadísticas generales de formas y lemas, 3) Estadísticas generales de lemas por L1, 4) Estadísticas generales de lemas por nivel, y 5) Frecuencias de lemas. Así, desde las muchas aristas posibles de comentar, baste la siguiente tabla que resume las estadísticas generales que sustentan la versión actual disponible.

Tabla 1. Estadísticas generales de formas y lemas

Total de formas registradas (tokens)	573.718
Total de formas distintas registradas (types)	38.655
Total de formas sin signos de puntuación	460.019
Total de formas sin signos de puntuación ni cifras ni fechas ni horas	458.475
Total de formas sin signos de puntuación ni cifras ni fechas ni horas ni nombres propios	438.828
Total de lemas registrados	16.643
Total de lemas sin signos de puntuación	16.382
Total de lemas sin signos de puntuación ni cifras ni fechas ni horas	15.505
Total de lemas sin signos de puntuación ni cifras ni fechas ni horas ni nombres propios	8.812

Como se observa, se dispone de rica información que permite visualizar solo algunos de los muchos campos a ser estudiados. Estos pormenorizados datos ilustran así las potencialidades del sitio, aún todavía sin haber ingresado a la aplicación misma de consulta.

Ahora bien, sin dejar de valorar las características del corpus y del diseño tecnológico en que se presenta el CAES, especial mención nos merece el documento "Guía de consulta de CAES" (disponible también en la página inicial). En nuestra opinión, dicho instrumento de trabajo constituye mucho más que una mera "guía de consulta", ya que brinda un amplio y diversificado instructivo de uso, al mismo tiempo que un muy sólido mapa de ruta de cómo moverse a través del sitio. Por ejemplo, se ofrece una explicación paso a paso del uso y explotación de las categorías de búsqueda y de su potencial empleo y aplicación, todo ello con ejemplos muy claros e ilustrativos. Cada apartado se acompaña de las correspondientes pantallas, tomadas del sitio, lo que resulta muy ilustrativo para quien tiene poco o nada de experiencia con principios de lingüística de corpus y desarrollos tecnológicos web de esta naturaleza. Destacamos particularmente este documento, pues muchos sitios que ofrecen corpus de aprendientes, en particular para el inglés como lengua extranjera, carecen de suficientes orientaciones y de un instructivo paso a paso como el que aquí comentamos.

Nos parece prudente sugerir que este documento pudiera mostrarse más ostensiblemente y destacarse de otro modo a como se presenta en la actualidad, de manera que se haga imprescindible acceder a él y su lectura y consulta sean así oportunas y productivas para los interesados. En su estado actual, es probable que muchos no alcancen a ver prontamente su tremendo potencial y parezca solo otro enlace a detalles de constitución del corpus o de estadísticas.

En la esquina izquierda de la pantalla inicial se encuentra el hipervínculo "Acceso a la aplicación de consulta". Mediante este se accede al eje central del CAES y se ingresa a la herramienta de búsquedas y consultas, tal como se puede apreciar por las casillas que aparecen en su página principal. Las consultas al CAES pueden hacerse de forma que se tengan en cuenta todos los parámetros utilizados en la configuración del corpus. Es decir, es posible realizar consultas en base a:

1. Nivel adquirido de conocimiento de español (de A1 a C1)
2. L1 (lengua inicial o familiar)
3. País de residencia
4. Edad
5. Sexo

A partir de estas casillas iniciales se puede alcanzar un alto grado de sofisticación en cada búsqueda posible, lo que no parece observable a simple vista. Es aquí justamente donde la "Guía de consulta de CAES" adquiere especial relevancia. Sin lugar a dudas, esta herramienta de consultas constituye en sí misma un tremendo avance desde el punto de vista tecnológico, así como un sustantivo soporte didáctico con un enorme potencial para la investigación empírica. Los pasos técnicos son reveladores de conocimientos teóricos y aplicados fundamentales para sustentar un desarrollo de esta naturaleza.

Junto a lo anterior, cabe señalar que todas las muestras textuales recolectadas y disponibles han sido anotadas morfosintácticamente de modo automático, mediante una versión modificada por el grupo de la Universidad de Santiago de Compostela de la aplicación FreeLing5. Estas marcas de etiquetaje luego fueron revisadas manualmente con el fin de alcanzar un alto grado de fiabilidad. Este segundo proceso reviste alta relevancia para contar con un corpus anotado correctamente. Gracias al marcaje estructural o anotación morfosintáctica que se ha aplicado al corpus, se posibilita una amplia diversidad de estudios tanto de tipo léxico como gramatical.

Entre las infinitas posibilidades de búsqueda que se ofrecen, hemos optado, a modo de ejemplo, por una consulta simple de concordancias para la ocurrencia del pronombre neutro "ello" en sujetos de nivel A1 de diversas lenguas de procedencia en países diversos:

Tabla 2. Ejemplos de una búsqueda de concordancia para el pronombre neutro "ello".

1 (A1/Francés)	, ha encontrado un hombre y se marza con	**ello**, se llama Jean-Yves, es un profesor de pianó, es símpatico
2 (A1/Portugués)	Yo no me parezco con	**ello**.
3 (A1/Portugués)	Extraño mucho	**ello**.
4 (A1/Árabe)	su marido que se llama JAMAL y trabajo con	**ello**.
5 (A1/Chino mandarín)	También	**ello** estudia español en Instituto_Cervantes en pekín .
6 (A1/Portugués)	mayor hace comunicación en la universidade_de_Brasilia, porque a	**ello** le gusta ser un periodista .
7 (A1/Portugués)	Mi padre es arquitecto, si llama Ricardo y	**ello** si que había nascido en São_Paulo.
8 (A1/Portugués)		**Ello** es profesor de curso superior de letras en la Universidad_de_Porto.
9 (A1/Portugués)		**Ello** és casado con Cláudia_de_Oliveira y ella tiene una hija llamada Caroline.
10 (A1/Portugués)		**Ello** vive en la Ciudad de Porto, en Portugal.

Como se indicaba, resulta muy valiosa la posibilidad de observar al mismo tiempo resultados de una búsqueda de materiales producidos por sujetos de distintas lenguas, en este caso francés, portugués, árabe y chino mandarín. Como se aprecia en los ejemplos, el pronombre neutro "ello" encierra en sí mismo una diversidad de opciones textuales para su uso, incluso muy complejas de

manejar para escritores de español como lengua materna. Más aún, sus funcionalidades todavía están lejos de ser estudiadas en profundidad en tanto mecanismo encapsulador de soporte a la construcción de la coherencia textual, tanto en la comprensión como en la producción de textos escritos (Parodi y Burdiles 2015a). Asimismo, desde cada lengua de procedencia habrá que determinar en estudios contrastivos el valor de dicho pronombre, si este existe, y sus diversos usos como elemento de coherencia textual tanto referencial como relacional. En este sentido, como puede imaginarse, las potencialidades para la investigación básica y aplicada son infinitas, tal como se desprende de algunos de los casos ilustrados en la tabla anterior. Sería prudente, eso sí, insistir en la advertencia a los usuarios del sitio acerca de que la disponibilidad del corpus en su estado original implica que sus muestras incluyen, muchas veces, errores típicamente producidos por escritores aprendientes de español como lengua extranjera en una fase de su desarrollo lingüístico, influido por el tipo de lengua materna de origen.

Todo lo anterior redunda en un sitio que se hace accesible a investigadores y, por qué no, a estudiantes, y se potencia así aún más su utilidad. En este sentido, otra de las muchas fortalezas que la iniciativa CAES brinda es la de ser uno de los primeros sitios web disponibles en línea de libre acceso para los interesados en la investigación del español como lengua extranjera, en el cual se definen muy claramente los parámetros y focos de las búsquedas.

Sin negar la relevancia que muestran otros proyectos de corpus de este tipo en el mundo, el CAES avanza de modo innovador para llenar un vacío sustancial en el área específica de corpus de aprendientes de español. Ello lo hace de modo similar a la amplia y prolífica gama de desarrollos para inglés como lengua extranjera (al respecto, ver el apartado “Learner corpora around the world” en la página de la LCA), tal como se puede ejemplificar con el desarrollo del ICLE (*International Corpus of Learner English*) en la Universidad de Lovaina. En esta línea, cabe también destacar los avances del equipo coordinado por Anita Ferreira, quien desde Chile lleva a cabo investigaciones y desarrollos tecnológicos con corpus de aprendientes en torno al creciente impacto del español como lengua extranjera (Ferreira, Vine y Ejalde 2013, 2014). Del mismo modo, también resulta ilustrativo destacar el trabajo de Rodgers, Chambers y Le Baron-Earle (2011) con un desarrollo de corpus para el francés, particularmente en el ámbito de la biotecnología.

En síntesis, el sitio web que analizamos constituye un aporte muy robusto al área, no solo por la disponibilidad de contar con un corpus diversificado en un conjunto de variables que posibilita la investigación para el español como lengua extranjera, sino también por disponer de un diseño con soportes cuidadosamente pensados y elaborados de modo muy profesional y sustentados en documentación complementaria altamente útil. El equipo de la Universidad de Santiago de Compostela con apoyo del Instituto Cervantes ha realizado un trabajo científico y tecnológico de gran calidad. Ahora bien, a modo de recomendación final, solo señalamos que sería oportuno revisar el nombre del corpus, ya que se define con foco en “el español como lengua extranjera”, pero esto no se recupera en el nombre genérico del título ni en su acrónimo “corpus de aprendices del español (CAES)”, lo que podría, eventualmente, conducir a algún tipo de error de comprensión inicial.

Por último, entre los muchos desafíos para el área, podemos reiterar lo propuesto por Parodi (2010b) en cuanto a avanzar más allá del estudio de corpus que atiendan a solo uno de los sistemas constitutivos de un texto, como regularmente se hace, esto es, únicamente al verbal. Los rasgos multisemióticos o multimodales esenciales en los textos de géneros discursivos especializados, tales como gráficos, esquemas, tablas, figuras, ilustraciones, se constituyen en un reto que debe superar el logocentrismo imperante (Parodi 2010c, 2010d, 2015, Parodi y Burdiles 2015b). Así, siguiendo la idea de Johns (1986, 2), sería oportuno considerar que “If computers

are to act as informants, the problem is how to get the machine to respond to learner-generated questions. The obvious answer is that we should try to make it as intelligent as possible…".

Bibliografía

Aijmer, K., ed. 2009. *Corpora and Language Teaching.* Amsterdam: Benjamins.

Baker, P., ed. 2009. *Contemporary Corpus Linguistics*. London: Continuum.

Facchinetti, R., ed. 2007. *Corpus Linguistics 25 Years on*. Amsterdam: Rodopi.

Ferreira, A., A. Vine y J. Elejalde. 2013. "Hacia una prueba de nivel en español como lengua extranjera". *Revista de lingüística teórica y aplicada* 51 (2): 73–103.

Ferreira, A., A. Vine y J. Elejalde. 2014. "Análisis de Errores Asistido por Computador basado en un Corpus de Aprendientes de Español como Lengua Extranjera". *Revista Signos. Estudios de Lingüística* 47 (86): 385–411.

Flowerdew, L. 2015. "Data-driven Learning and Language Learning Theories: Whither the Twain Shall Meet". En *Multiple Affordances of Language Corpora for Data-driven Learning*, eds. A. Lenko-Szymanska y A. Boulton, 15–36. Amsterdam: John Benjamins.

Gavioli, L. 2005. *Exploring Corpora for ESP Learning.* Amsterdam: John Benjamins.

Granger, S., G. Gilquin y F. Meunier, eds. 2013. *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S., J. Hung y S. Petch-Tyson, eds. 2002. *Computer Learner Corpora. Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Johns, T. 1986. "Micro-Concord: A Language Learner's Research Tool". *System* 14 (2): 151–62.

McEnery, T. y A. Hardie, eds. 2011. *Corpus Linguistics: Method, Theory and Practice.* Cambridge: Cambridge University Press.

Meunier, F., S. De Cock, G. Gilquin y M. Paquot, eds. 2011. *A Taste for Corpora. In Honour of Sylviane Granger*. Amsterdam: John Benjamins.

Parodi, G. 2010a. *Lingüística de corpus: de la teoría a la empiria*. Frankfurt: Editorial Iberoamericana-Veruert.

Parodi, G. 2010b. "Research Challenges for Corpus Cross-Linguistics and Multimodal Texts". *Information Design Journal* 18 (1): 69–73.

Parodi, G. ed., 2010c. *Academic and Professional Discourse Genres in Spanish*. Amsterdam: John Benjamins.

Parodi, G. 2010d. "Multisemiosis y lingüística de corpus: artefactos (multi)semióticos en los textos de seis disciplinas en el corpus PUCV-2010". *Revista de lingüística teórica y aplicada* 48 (2): 33–70.

Parodi, G. 2015 (en prensa). "Variation across University Genres in Seven Disciplines: A Corpus-Based Study on Academic Written Spanish". *International Journal of Corpus Linguistics* 20 (4).

Parodi, G. y G. Burdiles. 2015a. "El pronombre 'ello' como mecanismo encapsulador en cuatro géneros del discurso de la Economía: coherencia referencial y relacional". En *Leer y escribir en contextos académicos y profesionales: géneros, corpus y métodos*, eds. G. Parodi y G. Burdiles, 445–82. Santiago de Chile: Ariel.

Parodi, G. y G. Burdiles, eds. 2015b. *Leer y escribir en contextos académicos y profesionales: géneros, corpus y métodos*. Santiago de Chile: Ariel.

Rodgers, O., A. Chambers y F. Le Baron-Earle. 2011. “Corpora in the LSP Classroom: A Learner-Centred Corpus of French for Biotechnologists”. *International Journal of Corpus Linguistics* 16 (3): 391–411.
Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Sinclair, J. 1994. “Trust the Text”. En *Advances in Written Text Analysis*, ed. M. Coulthard, 12–25. London: Routledge.
Timmins, I. 2015. *Corpus Linguistics for ELT: Research and Practice*. London: Routledge.

Giovanni Parodi
Pontificia Universidad Católica de Valparaíso
gparodi@ucv.cl