"HARTA" de noveles
Un corpus de español académico

Milka Villayandre Llamazares
Universidad de León

The aim of this review is to account for the process of compilation and codification of the corpus HARTA-Noveles. This corpus was created as part of the research project titled "Corpus-based study of lexical combinations of academic Spanish for the development of a computational tool for academic writing assistance" (HARTA)[1], under the direction of Margarita Alonso Ramos (University of La Coruña). The corpus consists of representative samples of essays produced by Spanish university students and gathered with the purpose of studying academic lexical combinations (CLA)[2], i.e., recurrent segments specific to the academic domain, along with collocations, discourse markers and other multiword expressions. Inspired by the BAWE corpus (British Academic Written English), our corpus is formed exclusively by final project texts (for the degrees) and dissertations (for the masters) selected from different public repositories of Spanish universities and from various scientific domains. These texts have been annotated with an specific system adapted from that followed by the Spanish Royal Academy (RAE) in CORPES XXI[3].

Keywords: corpus linguistics, learner corpus, academic writing, lexical combinations, writing assistant

1. Introducción

Si bien los corpus de aprendiz son un campo joven dentro de la lingüística de corpus, en el caso del español hay que decir que estamos asistiendo aún a sus primeros pasos. M. Alonso-Ramos (2016) recoge en un volumen monográfico las principales investigaciones y proyectos que se han llevado o se están llevando a

[1] Acrónimo del nombre en español del proyecto: "Estudio de las combinaciones léxicas del español académico basado en corpus para una Herramienta de Ayuda a la Redacción de Textos Académicos" (HARTA).
[2] En español, combinaciones léxicas académicas (CLA).
[3] Corpus del Español del Siglo XXI (CORPES).

CHIMERA. Romance Corpora and Linguistic Studies 5.1 (2018), 131-140. ISSN 2386-2629

cabo en esta línea de trabajo. Si, además, se busca un corpus centrado en escritos académicos, el abanico de posibilidades se reduce todavía más. Cabe citar el Corpus académico y profesional de la Pontifica Universidad Católica de Valparaíso, en Chile (Parodi 2007, 2009) o el Corpus de lenguaje académico en español (CLAE 2009), de la Universidad Nacional Autónoma de México (UNAM) y la Universidad de California, Davis (UCD). A estos corpus hay que sumar la celebración en fechas recientes del I Congreso Internacional sobre Análisis de Corpus del Discurso Académico, que tuvo lugar en Valencia, del 22 al 24 de noviembre de 2017[4].

En este contexto surge el corpus HARTA-Noveles, parte del proyecto de investigación "Estudio de las combinaciones léxicas del español académico basado en corpus para una Herramienta de Ayuda a la Redacción de Textos Académicos (HARTA), dirigido por Margarita Alonso Ramos (Universidad de La Coruña) y financiado por el Ministerio de Economía y Competitividad.

El corpus, que en el momento actual se ha terminado de compilar, pretende recoger muestras representativas de la escritura académica de estudiantes universitarios españoles con el objeto de estudiar posteriormente combinaciones léxicas académicas (CLA), es decir, segmentos recurrentes específicos del ámbito académico; junto a colocaciones, marcadores discursivos y otro tipo de fórmulas. Este es un paso previo para desarrollar una herramienta de ayuda a la redacción de escritos académicos, objetivo final de la investigación.

El proyecto toma su inspiración del corpus BAWE (British Academic Written English), de las Universidades de Warwick, Reading y Oxford Brookes, en el Reino Unido (Ebeling y Heuboeck 2007; Nesi 2008a, 2008b; Alsop y Nesi 2009), aunque se ha adaptado a las necesidades del proyecto. Por otra parte, introduce un nuevo matiz en el panorama de los corpus de aprendiz, pues los textos proceden de hablantes nativos y no de hablantes no nativos. Se trata de contrastar sus producciones con las del corpus HARTA-Expertos, conformado por artículos científicos de especialistas en las distintas materias.

El diseño y recopilación del corpus HARTA-Noveles se inició en mayo de 2017 y ha finalizado en agosto de 2018.

2. Tipo de corpus y parámetros de selección de textos

Como se ha mencionado en el apartado anterior, HARTA-Noveles es un corpus de aprendiz, compuesto únicamente por textos producidos por estudiantes en el

[4] Hay que señalar que la mayor parte de las contribuciones se centraron en otras lenguas, aunque la iniciativa da cuenta de lo reciente del interés suscitado por los corpus de textos académicos.

ámbito universitario, esto es, de tipología académica. En concreto, se trata de Trabajos de Fin de Grado (TFG) y Trabajos de Fin de Máster (TFM).

La extensión total del corpus es de algo más de 2.000.000 de palabras, distribuidas en cuatro hipercampos: Arte y humanidades, Biología y ciencias de la salud, Ciencias físicas y Ciencias sociales, la misma distribución que presenta el corpus BAWE. Cada uno contiene unas 500.000 palabras, 250.000 procedentes de TFG y otras 250.000 de TFM, de diferentes disciplinas, como se ilustra en las tablas 1[5] y 2.

Tabla 1. Hipercampos y áreas temáticas en HARTA-Noveles

Arte y human.	Biología y ciencias de la salud	Ciencias físicas	Ciencias sociales
Arte	Biología	Ciencias de la Tierra	Econ. y empresa
Lingüística	Medicina	Física	Educación
Literatura		Ingeniería	Sociología
Historia y cultura		Informática	
Biblioteconomía y documentación		Química	

Tabla 2. Distribución de palabras y textos por hipercampos y por tipología de trabajo (TFG o TFM)

Hipercampo	TFG	TFM	Total palabras (y textos)
Arte y humanidades	403 562 (25)	289 973 (13)	693 535 (38)
Biología y ciencias de la salud	258 618 (38)	250 567 (30)	509 185 (68)
Ciencias físicas	249 074 (21)	249 508 (21)	498 582 (42)
Ciencias sociales	275 081 (17)	253 770 (11)	528 851 (28)
Totales	**1 186 335 (101)**	**1 043 818 (75)**	**2 230 153 (176)**

5 Las materias o áreas temáticas están condicionadas por aquellas presentes en HARTA-Expertos, corpus con el que se van a contrastar los resultados, así como por la disponibilidad de materiales en los repositorios digitales empleados como fuente de los textos.

En relación a su procedencia, los textos que conforman el corpus se seleccionaron de forma aleatoria de los repositorios públicos de diferentes universidades españolas. En concreto, de las siguientes:

- Una parte muy importante del corpus está compuesta por textos de la Universidad de La Coruña (UDC), como entidad promotora del proyecto[6]. Estos suponen un 61,93% del total (109 textos).
- Otra parte de los textos procede de la Universidad de León (ULE), que también colabora en el proyecto[7]. Los textos de la ULE corresponden al 18,18% del total (32 textos).
- Finalmente se incorporaron textos de otras universidades para completar las áreas temáticas cuya representación era insuficiente en los repositorios de la UDC y de la ULE: Universidad de La Laguna (ULL, 3 textos), Universidad de Salamanca (USAL, 14 textos), Universidad Complutense de Madrid (UCM, 5 textos) y Universidad de Valladolid (UVA, 13 textos); en total, 35 textos que corresponden al 19,88% del total del corpus[8].

La distribución de textos por universidad es la siguiente (tablas 3 y 4):

Tabla 3. Distribución del número de textos por cada universidad

	UDC	**ULE**	**ULL**	**USAL**	**UCM**	**UVA**
Textos y %	109 61,93%	32 18,18%	3 1,70%	14 7,95%	5 2,84%	13 7,39%
Subtotales	141 textos (80,11%)		35 textos (19,88%)			
Total	**176 textos**					

Tabla 4. Distribución de palabras por universidad

	UDC	**ULE**	**ULL**	**USAL**	**UCM**	**UVA**
Número de palabras	1 126 267	570 457	36 698	231 643	77 216	187 872

[6] Repositorio RUC: https://ruc.udc.es/dspace/handle/2183/7189.
[7] Repositorio BULERIA: https://buleria.unileon.es/
[8] Repositorio institutcional de La Universidad de La Laguna: https://riull.ull.es/xmlui/handle/915/668; repositorio de la Universidad de Salamanca: https://gredos.usal.es/jspui/handle/10366/4829; repositorio de la Universidad Complutense de Madrid: https://eprints.ucm.es/information.html; repositorio de la Universidad de Valladolid: http://uvadoc.uva.es/handle/10324/38.

Subtotales	50,50%	25,58%	1,65%	10,39%	3,46%	8,42%
Total	**2 230 153 palabras**					

3. Sistema de codificación del corpus

El sistema de codificación del corpus HARTA-Noveles se basa en el empleado en el Corpus del español del siglo XXI (CORPES) de la Real Academia Española (Real Academia Española 2013).

Se han incluido muchos elementos de este sistema de codificación con la intención de que la extracción de datos del corpus HARTA sea similar e, incluso, de que ambos corpus se puedan compatibilizar en el futuro para desarrollar investigaciones en el ámbito de la lingüística, pese a que uno es de carácter especializado (HARTA) y el otro, general o de referencia (CORPES). No obstante, esta diferencia ha sido determinante para la adaptación del esquema de codificación, que es la manera de identificar y codificar cada texto del corpus.

Un codificador con experiencia previa en CORPES ha sido el responsable de llevar a cabo el proceso de diseño y validación del esquema de codificación, así como la compilación del mismo, con la supervisión de la investigadora principal y el asesoramiento del resto de miembros del equipo de investigación.

El procedimiento seguido para la codificación de los textos se describe a continuación:

i. En primer lugar, se han descargado de forma aleatoria los textos de los repositorios digitales de las distintas universidades, que estaban en formato PDF.
ii. En segundo lugar, los textos originales en PDF se han convertido a HTML por medio del programa Abby PDF Transformer 3.0[9], con el fin de no perder la estructura del texto en párrafos y otras marcas de formato relevantes para la extracción de información.
iii. En tercer lugar, los documentos .html se han convertido a documentos de texto (TXT). De esta manera ya se puede trabajar con ellos mediante un editor de textos.
iv. El siguiente paso ha consistido en la codificación propiamente dicha del texto. Mediante una lectura superficial, se procede a la limpieza del mismo, con la eliminación de las marcas de formato y de otro tipo que no interesan y se introducen las que están definidas en el sistema de

[9] URL: https://www.abbyy.com/en-ca/support/pdftransformer/30/sr/

codificación. Esta operación se realiza con un editor de texto, TextPad®[10].

v. Por último, ya como documentos XML, se les ha añadido la cabecera a los textos y estos son validados con el programa Oxygen[11]. Este es un potente editor de XML que permite asegurar el uso correcto de las etiquetas definidas en el esquema, así como la recuperación posterior de la información.

Las etiquetas empleadas en el sistema de codificación son las siguientes:

a. Etiquetas comunes con el sistema de codificación de la RAE. Estas ya están presentes en CORPES y sirven a los propósitos de HARTA-Noveles. Se trata de las siguientes (tabla 5):

Tabla 5. Etiquetas comunes con el sistema de codificación de la RAE para CORPES

Etiqueta	Descripción
…	Marca de párrafo
_…	El texto comprendido entre estas etiquetas está subrayado.
<csv>…</csv>	Indican que el texto está en cursiva.
<ngr>…</ngr>	Señalan que el texto está en negrita.
<vrs>…</vrs>	Indican que el texto está en versales.
<csvngr>…</csvngr>	El texto está al mismo tiempo en cursiva y en negrita.
<sic>…</sic>	Indica un fragmento de texto confuso, que no se entiende.
<rsi>…</rsi>	Texto resaltado de manera diferente.
<nrp/>	Fragmento no representable, eliminado, sin contenido.

b. Etiquetas creadas para HARTA-Noveles. Fue necesario introducir etiquetas nuevas para marcar aspectos estructurales habituales en los textos académicos y, por tanto, relevantes para la finalidad del proyecto, como se especifica en la tabla 6.

Tabla 6. Nuevas etiquetas adoptadas para el corpus HARTA-Noveles

Etiqueta	Descripción
<nrp>…</nrp>	Comprende títulos de tablas y figuras que se han eliminado, así como su pie, si lo tuvieran.

[10] URL: https://www.textpad.com/

[11] URL: https://www.oxygenxml.com/

<idm>…</idm>	Indica texto en un idioma extranjero.
<cita>…</cita>	Comprende texto citado que aparece como párrafo independiente.
<titl>…</titl>	Títulos y subtítulos de apartados del trabajo.

Todos los textos que conforman el corpus cuentan con un identificador único para cada uno ("id"), de manera que sea más fácil trabajar con los mismos. El "id" consta de dos letras mayúsculas que remiten al hipercampo y tres dígitos que identifican el texto. Por ejemplo, el "id" CS_018 nos indica que estamos ante el texto número 18 de Ciencias Sociales. Los códigos para los hipercampos se detallan a continuación:

- AH: Arte y Humanidades
- BC: Biología y Ciencias de la Salud
- CF: Ciencias Físicas
- CS: Ciencias Sociales

La estructura de cada texto consta de tres partes: i) información sobre el esquema de codificación e "id"; ii) cabecera; iii) contenido o texto propiamente dicho.

La "cabecera" contiene todos los datos necesarios para identificar el texto y recuperar la información. En concreto, contiene los siguientes campos o elementos:

- El campo "trabajo" sirve para definir el título y el autor:
  <trabajo título="" autor=""></trabajo>
- El campo "depósito" define la procedencia del trabajo y el año de su elaboración (año de depósito). El atributo "universidad" tomará uno de los posibles valores establecidos: Universidad Complutense de Madrid, Universidad de La Coruña, Universidad de La Laguna, Universidad de León, Universidad de Salamanca o Universidad de Valladolid.
  <depósito año_depósito="" universidad=""/>
- El campo "clasificación" describe el tipo de trabajo. El atributo "tipología" siempre toma el valor de "académico", ya que no hay otro tipo de textos en este corpus. El atributo "tipo de trabajo" puede tomar dos valores: TFG o TFM. El atributo "tema" puede ser uno de los hipercampos en los que se ha dividido el corpus: Arte y humanidades, Biología y ciencias de la salud, Ciencias físicas o Ciencias sociales. El atributo "subtema" toma cualquiera de los valores resumidos en la tabla 1.
  <clasificación tipología="académico" tipo_trabajo="" tema="" subtema=""/>

- El campo “extensión” nos informa del número de palabras del texto:
  <extensión palabras=""/>

Todos estos atributos son obligatorios para que el editor de XML valide el documento, pues así se ha definido en el esquema desarrollado para la codificación de HARTA-Noveles.

Dentro del elemento “contenido” se recoge el texto con las etiquetas establecidas en el sistema de codificación (tablas 5 y 6). Cada texto puede estar estructurado de manera diferente. Por eso, se ha diseñado un esquema flexible que permite utilizar algunos elementos de manera opcional, según el texto concreto. Lo mínimo que tendrá el texto será un “cuerpo”, pero el máximo de apartados puede ser el siguiente:

- Resumen
- Presentación
- Introducción
- Cuerpo. Opcionalmente puede contener otros tres elementos:
  - Métodos
  - Resultados
  - Discusión
- Conclusión
- Agradecimientos
- Notas
- Anexos (por si se considera oportuno incluir alguno)

Este esquema da cuenta de la estructura de los diferentes trabajos que forman parte del corpus.

4. Conclusión y resumen del esquema de codificación

Lo aquí expuesto resume el trabajo llevado a cabo hasta la fecha para el diseño y compilación del corpus HARTA-Noveles.

Actualmente se está procediendo a su revisión, así como al etiquetado automático, por lo que el corpus final puede diferir ligeramente del aquí presentado, en la medida en que sea necesario efectuar algún ajuste.

Por lo que respecta al esquema y sistema de codificación, partiendo de la experiencia de parte del equipo de investigación en CORPES, se han mantenido aquellos procedimientos y etiquetas que resultaban relevantes para los fines del

trabajo. En los casos en los que el tipo de textos exigía contar con nuevas marcas que permitieran identificar mejor los elementos estructurales propios del discurso académico, se introdujeron etiquetas adicionales.

Para finalizar, se presenta un resumen del esquema de codificación establecido para el corpus HARTA-Noveles:

```
<?xml version="1.0" encoding="UTF-8"?>
<HARTA xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="file:/C:/UDC/harta.xsd" id="">
 <cabecera>
   <trabajo título="" autor=""></trabajo>
   <depósito año_depósito="" universidad=""/>
   <clasificación tipología="académico" tipo_trabajo="" tema="" subtema=""/>
   <extensión palabras=""/>
 </cabecera>
 <contenido>
   <resumen>
     <p></p>
   </resumen>
   <presentación>
     <p></p>
   </presentación>
   <introducción>
     <p></p>
   </introducción>
   <cuerpo>
     <p></p>
     <métodos>
       <p></p>
     </métodos>
     <resultados>
       <p></p>
     </resultados>
     <discusión>
       <p></p>
     </discusión>
   </cuerpo>
   <conclusión>
     <p></p>
```

```
    </conclusión>
    <agradecimientos>
      <p></p>
    </agradecimientos>
    <notas>
      <p></p>
    </notas>
    <anexos>
      <p></p>
    </anexos>
    </contenido>
  </HARTA>
```

Referencias bibliográficas

Alonso Ramos, M. (ed.) 2016. *Spanish Learner Corpus Research: Current trends and future perspectives*. Amsterdam: John Benjamins.

Alsop, S. y Nesi, H. 2009. Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora* 4 (1) 71-83.

CLAE: *Corpus de lenguaje académico en español*. 2009. <http://www.lenguajeacademico.info>. UCMexus-CONACYT (Consultado el 1 de junio de 2018).

Ebeling, S and Heuboeck, A. 2007. Encoding document information in a corpus of student writing: the experience of the British Academic Written English (BAWE) corpus. Encoding document. Corpora 2 (2) 241-256

Nesi, H. 2008a. BAWE: An introduction to a new resource. En Frankenberg-Garcia, A., Rkibi, T., Braga da Cruz, M., Carvalho, R., Direito, C. & Santos-Rosa, D. (eds) Proceedings of the 8th Teaching and Language Corpora Conference. Held 4-6 July 2008 at the Instituto Superior de Línguas e Administração,. Lisbon, Portugal: ISLA: 239-246.

Nesi, H. 2008b. 'Introducing BAWE: a new lexicographical resource.' En Bernal, E. and DeCesaris, J. (eds.) Proceedings of the XII EURALEX International Congress. Held 15-19 July 2008 at Universitat Pompeu Fabra. Barcelona: Institut Universari de Linguistica Aplicada, Universitat Pompeu Fabra: 737-752.

Parodi, G., Ed. 2007. *Lingüística de corpus y discursos especializados: Puntos de Mira*. Valparaíso: Ediciones Universitarias de Valparaíso.

Parodi, G. 2009. El corpus académico y profesional del español PUCV-2006: semejanzas y diferencias entre los géneros académicos y profesionales. *Estudios Filológicos* 44, 123-147.

Real Academia Española. 2013. *Corpus del español del siglo XXI (CORPES). Descripción del sistema de codificación. Libros y prensa*. Madrid: Real Academia Española.