Experimenta: Revista de divulgación científica de la Universidad de Antioquia, vol. 9, 2018, pp. 52-55.

Corpus Ex Machina: Computación y lengua española.

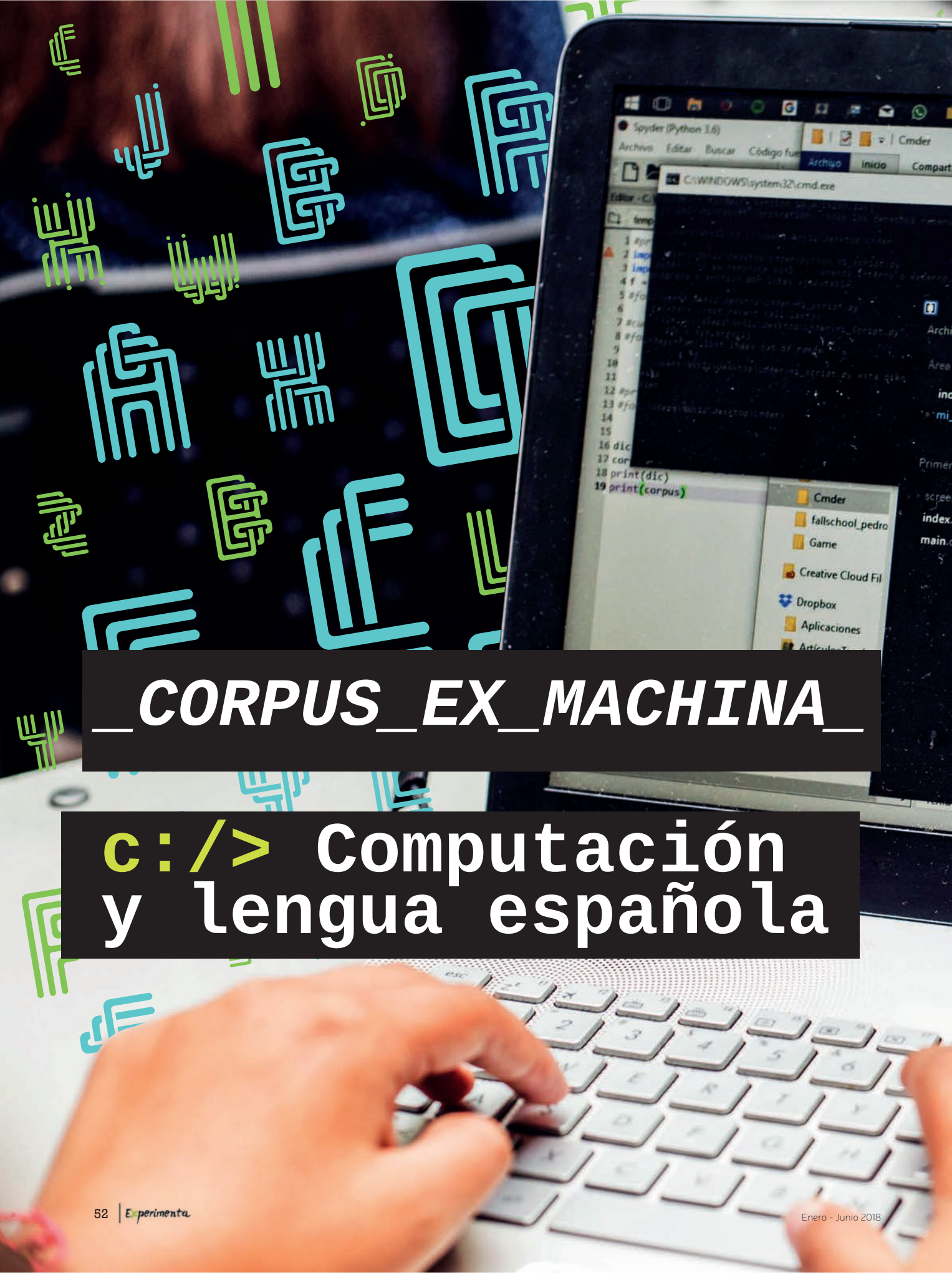
Molina Mejia, Jorge Mauricio, Grajales Ramírez, Andrés Felipe, Pemberty Tamayo, José Luis y Zapata Granados, Maria Adelaida.

Cita:

Molina Mejia, Jorge Mauricio, Grajales Ramírez, Andrés Felipe, Pemberty Tamayo, José Luis y Zapata Granados, Maria Adelaida (2018). *Corpus Ex Machina: Computación y lengua española. Experimenta: Revista de divulgación científica de la Universidad de Antioquia, 9, 52-55.*

Dirección estable: https://www.aacademica.org/jorge.mauricio.molina.mejia/45

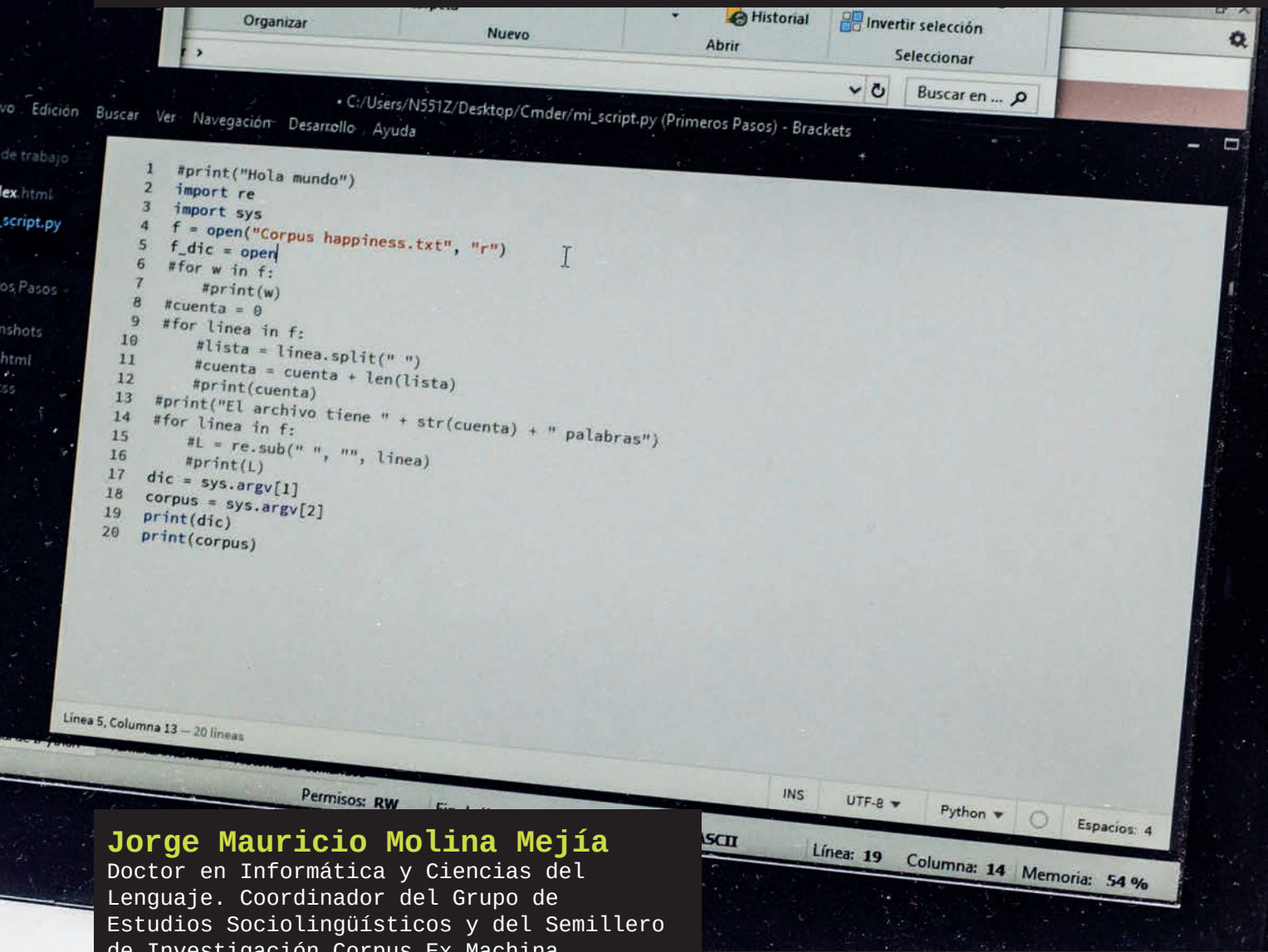
ARK: https://n2t.net/ark:/13683/pqc6/2gf



_CORPUS_EX_MACHINA_

c:/> Computación y lengua española

Corpus Ex Machina es un semillero de investigación en las áreas de la Lingüística Computacional y de la Lingüística de Corpus que se encarga del estudio del lenguage natural y, para ello, propone el empleo de sistemas informáticos que posibilitan su análisis y aprendizaje.



Jorge Mauricio Molina Mejía
Doctor en Informática y Ciencias del Lenguaje. Coordinador del Grupo de Estudios Sociolingüísticos y del Semillero de Investigación Corpus Ex Machina.

Andrés Felipe Grajales Ramírez
Estudiante del pregrado en Letras: Filología Hispánica.

José Luis Pemberty Tamayo
Estudiante del pregrado en Letras: Filología Hispánica.

Adelaida Zapata Granados
Estudiante del pregrado en Licenciatura en Literatura y Lengua Castellana.

El estudio del lenguaje es, sin lugar a duda, un campo transversal a todas las demás disciplinas de las ciencias humanas (filosofía, historia, sociología, psicología, etc.) y también a las ciencias exactas, de la salud y la ingeniería. A pesar de que los estudios acerca del lenguaje han sido un poco relegados, podemos considerar que son de suma importancia, por cuanto este impregna todo lo que se hace en las áreas antes mencionadas.

Por lo tanto, investigar en el área del lenguaje se ha convertido en algo que puede servir a estas otras disciplinas, como cuando analizamos el discurso científico y los artículos académicos.

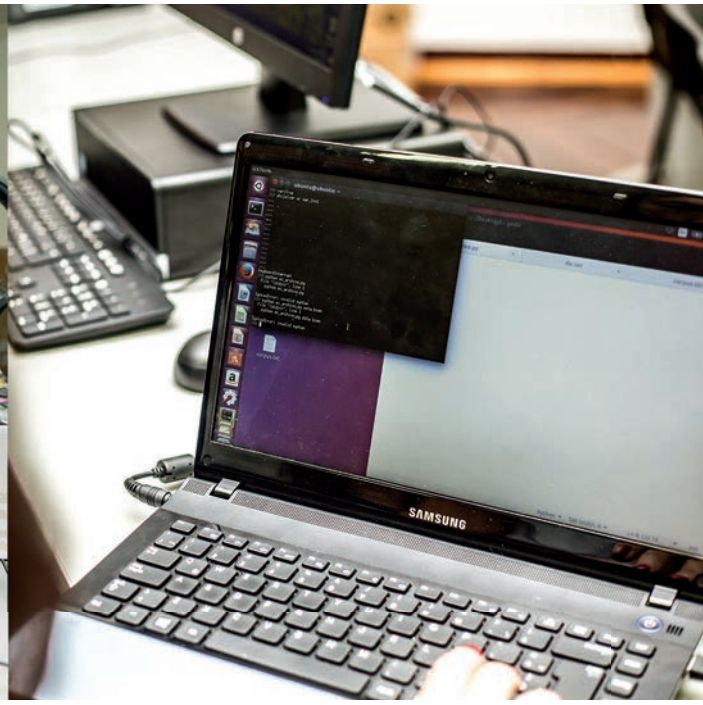
Corpus Ex Machina es un semillero de investigación en el área de la Lingüística Computacional adscrito al Grupo de Estudios Sociolingüísticos (Facultad de Comunicaciones) y a la Red de Semilleros de la Universidad de Antioquia (RedSin), que tiene por objeto el estudio del lenguaje a partir de las diferentes alternativas que ofrece la informática. En este sentido, los objetos de estudio o líneas de investigación que trabaja *Corpus Ex Machina* son la lingüística de *corpus*, la enseñanza de lenguas asistida por ordenador (ALAO) y el análisis de textos utilizando herramientas informáticas.

¿Qué es la lingüística computacional?

Podemos llamar *lingüística computacional* al campo que tiene por objeto el estudio del lenguaje a través de la computación. Permite comprender aplicaciones de esta disciplina en una gama de tareas tan amplia como la que representa el mismo uso del lenguaje en nuestra sociedad; por ejemplo, la traducción automática, el aprendizaje de lenguas, el mercadeo o el análisis artificial de textos y discursos. A su vez, este campo se encuentra atravesado por otras disciplinas que le permiten llevar a cabo sus diferentes tareas, como, por ejemplo: la inteligencia artificial, la lingüística teórica y aplicada, las ciencias cognitivas, las ciencias informáticas y de la programación, entre otros.

Para las diferentes aplicaciones de la lingüística computacional, a menudo es necesario contar con una gran cantidad de datos que son tomados como muestras de la utilización real de la lengua y que permiten llevar a cabo actividades tanto de análisis y comprensión de la manera en que nos comunicamos, como de realización de herramientas que faciliten la manera en la que lo hacemos. A esta gran cantidad de datos lingüísticos reunidos con un objetivo específico, se le llama *corpus lingüístico.*

En este sentido, ha sido necesario desarrollar un enfoque metodológico que resuelva las necesidades



de la recolección, el manejo y el procesamiento de estas amplias cantidades de datos en pro de mejorar los procesos de todas las ramas de la investigación de la lengua; es lo que llamamos *lingüística de corpus.*

¿Cómo aplicamos estos conceptos en el semillero?
Actualmente tenemos en desarrollo dos proyectos que utilizan todos los conceptos que ya hemos mencionado:

El Corpus PRESEEA-Medellín es un conjunto de entrevistas orales que se realizaron en 2005 en Medellín como parte del proyecto PRESEEA (Proyecto para el Estudio Sociolingüístico del Español de España y América) para estudiar y comprender la manera en que los habitantes de la ciudad utilizan la lengua, dependiendo de su edad, sexo, nivel educativo y estrato socioeconómico. Estas entrevistas se encuentran actualmente transcritas como textos y son utilizadas por el Grupo de Estudios Sociolingüísticos para diferentes investigaciones, además de ser accesibles al público en general.

El etiquetado de un corpus consiste en la utilización de un código específico para darle a los elementos de cada texto unas características definidas según el objetivo que se desee. Por ejemplo, en el caso del etiquetado morfosintáctico, que es el que se aplica al corpus PRESEEA-Medellín, se utiliza el lenguaje para marcar el comportamiento de cada palabra en cuanto a su morfología (si es de género masculino o femenino, si está en plural o en singular, etc.) y su papel en la sintaxis de las oraciones (si es un sujeto, un complemento, un verbo, etc.). Dicho de otro modo, el etiquetado consiste en traducir aquello que nosotros sabemos de las palabras y los textos a un lenguaje que el computador también pueda comprender para ser posteriormente utilizado de diferentes maneras.

En el caso de este proyecto, el objetivo es tener un etiquetado morfosintáctico de todos los textos para posteriormente implementar una plataforma en línea que permita estudiar la manera en la que las personas de Medellín utilizan la lengua, las palabras y las formas que usan, y también los rasgos particulares del habla de nuestra ciudad.

El Dispositivo informático para la enseñanza del español como lengua extranjera, DICEELE, es el otro proyecto que desarrolla el semillero y consiste en la utilización de la lingüística de corpus, la lingüística computacional y el aprendizaje de lenguas asistido por ordenador para ofrecer un *software* que se base en la utilización de varios textos clasificados según su nivel de dificultad, para ofrecer una interfaz que pueda ser usada por docentes para diseñar sus actividades, y por aprendientes para llevar a cabo procesos autodidactas. De esta manera tendríamos un sistema que permita tanto su aplicación en el aula de clase, como en la educación a distancia y en los procesos individuales. Estos textos también estarán etiquetados, pero esta vez con las nociones necesarias para comprender la composición de un texto en español.

La búsqueda principal de este proyecto es ofrecer una alternativa a la enseñanza y el aprendizaje tradicional de la lengua española, tanto para los extranjeros que vienen al país, como para las personas de comunidades dentro del mismo que no tienen el español como lengua materna.

El semillero utiliza la lingüística de corpus, la lingüística computacional y el aprendizaje de lenguas asistido por ordenador para ofrecer una interfaz que pueda ser usada por docentes para diseñar sus actividades, y por aprendientes para llevar a cabo procesos autodidactas