Tarea 5

Garduño Martínez Eduardo

Instrucciones: Lee el capítulo 2 .Accessing and analysing corpus data"de McEnery Hardie (2012). Con base en la presentación y la lectura del capítulo tal contesta el siguiente cuestionario. Adjunta un reporte con tus respuestas.

1. ¿Qué son los corpus y bajo qué criterios se pueden clasificar?

   Un corpus es un conjunto de textos, ya sea escritos o hablados, recolectados debidamente para realizar análisis lingüísticos, análisis que pueden ser cualitativos (estructurales) o cuantitativos (estadísticos). Pueden clasificarse respecto a sus características (por ejemplo, textuales u orales) o los elementos que poseen (por ejemplo, literarios o informativos).

2. Visita el sitio de un corpus de tu elección de los listados en la sección 4.5 y, siguiendo la tabla de la sección 4.1.2 de la presentación, caracterízalo.

   Corpus Científico del Español de México (COCIEM):

   *Origen de los datos: Textual.

   *Especificidad de los elementos: Corpus específico: Informativos: Científicos.

3. ¿Cuáles son los tres tipos de información que contiene un corpus además de los textos o transcripciones? Descríbelos brevemente.

   Metadatos: Información sobre el texto mismo, como idioma o fecha de publicación.

   Marcado textual: Codificación de la información que no provenga de palabras, como el momento en que un orador habla en una transcripción, o las itálicas o negritas en un texto.

   Notación lingüística: Codificación sistemática o manual para pronta recuperación de datos reelevantes, previamente observados en el texto o transcripción.

4. ¿Por qué resulta tan importante la consistencia en la anotación de un corpus?

   Debido a la confianza que se le brinda a una investigación lingüística que tiene consistencia de anotaciones, así como a la capacidad de comparar o intentar comparar con una investigación así con otras investigaciones.

5. ¿Cuáles son las principales características de las cuatro generaciones de concordancers?

   Primera generación: Actuaban en un mainframe de computadora que podía generalmente solo ejecutarse en esa computadora, estaban limitados a un solo acuerdo (el KWIC), y tenían bastante problemática al no tratar con caracteres romanos no acentuados.

   Segunda generación: Permitieron la diversificación en computadoras personales (como las de IMB). Sin embargo, la poca potencia de las computadoras del momento dieron desventajas inclusive contra la primera generación, teniendo problemas de identificación de cadenas o de escala de cantidad de información a procesar.

   Tercera generación: Revolucionaron la manera de analizar corpus, al manejar paquetes de diversas herramientas (como poder anotaciones en el corpus) y acuerdos, para procesar grandes cantidades de información en las PC's, siendo que eran capaces también de comprender varios sistemas de escritura, así como de realizar análisis estadísticos.

   Cuarta generación: Representan una evolución a la época moderna de la tercera generación, donde pueden adaptarse a trabajar con administradores de bases de datos o arquitecturas de software o

de trabajo modernas. Y aunque no representan un enorme cambio respecto a la tercera generación, son mucho más fáciles de utilizar y tienen más potencia, además resuelven problemas específicos que están relacionados con la limitación en algunas PC's o con la legalidad del compartimiento de corpus entre diferentes investigadores.

6. ¿Cuáles son las tres herramientas más usadas en el análisis de corpus y qué utilidad tienen?

   Concordancer: Busca cadenas con un cierto número de caracteres.

   Anotador de corpus: Realiza anotaciones sistemáticas.

   Analizador estadístico: Realiza conclusiones según la cantidad de datos existentes.

7. McEnery Hardy mencionan una idea alternativa a usar software open-source para el análisis de corpus. ¿Cuál es esta y cuáles son sus ventajas?

   En general, la idea consiste en que que haya softwares que puedan ser editados por varias personas de manera sencilla y remota. De forma que, como ventaja se tendría un avance colectivo sin necesidad reinventar la rueda"(rediseñar y/o reescribir código para cierta función) tan constantemente, enfocando ese tiempo en resolver los verdaderos problemas que se hayan actualmente en el análisis de corpus.

8. ¿Cuáles son las desventajas que tiene esta alternativa?

   Como algunos programas de búsqueda y recuperación de corpus son sistemas enormemente complicados, a veces, solo siendo comprensibles en su totalidad por su propio creador, resulta bastante complicado hacer totalmente viable un entorno donde todos los investigadores puedan desarrollarse con claridad y puedan, realmente ahorrar tiempo y utilizarlo en los problemas actuales.

9. ¿Qué es la frecuencia relativa de una palabra en un corpus y cómo se calcula?

   La frecuencia relativa de una palabra se calcula de la siguiente manera:

   fr = (numero de veces que aparece la palabra ÷ totalidad del corpus) × (base de normalización)

   Puede interpretarse como "¿que tan seguido podremos ver una palabra w en una cantidad n de palabras del corpus?"

10. ¿Cuáles son los principales tests de significatividad y por qué resultan relevantes?

    Los dos principalmente usados son calculating keywords y calculating colocations. Son importantes porque respectivamente, catalogan ciertas palabras como más significantes que otras según su cantidad de apariciones en un corpus con respecto a otro corpus de referencia y correlacionan la aparición de ciertas palabras (si una palabra aparece, que tan frecuente aparece otra cerca de ella).