0.1. Corpus

Un corpus es una colección de textos escritos o transcripciones lenguaje oral para cierto idioma, generalmente conformado por más de un millón de palabras. La palabra corpus viene del latin...., su plural es corpora. Corpus linguístico es el estudio del lenguaje como está expresado en los ejemplos. Este método representa una aproximación a derivar un conjunto de reglas abstractas con las cuales un lenguaje natural es gobernado. Este proceso originalmente fué realizado a mano, corpora ahora son ampliamente derivados por un autómata.

Diferentes tipos de de corpora permiten el análisis de distintas clases de discuross para hallar evidencia cuantitativa sobre la existencia de patrones en el lenguaje o para verificar teorías. Los primeros estudios sobre un corpus se enfocaron en palabras; su frecuencia y ocurencia. Con el desarrollo de la tecnología y de motores de búsqueda más precisos, las posibilidades crecieron ampliamente. Ahora es posible realizar búsquedas para una palabra perteneciente a cierta clase sintáctica o patrones completos como por ejemplo: preposición + sustantivo o determinante + sustantivo, o una palabra particular + clase de palabra específica sucediéndola. Estas búsquedas facilitan, por ejemplo, a los publicadores de diccionarios para hallar colocaciones.

La linguistica también es aplicada sobre el corpus para estudios de traducción donde utilizando corpora de 2 lenguajes distintos se hacen aparentes los significados de las palabras y sus supuestas equivalencias que podrían diferir en su uso o colocaciones. Más aún, algunos aspectos gramaticales fuertemente conectados a la léxica permiten a los linguistas mostrar diferencias en el uso de ciertas estructuras gramaticales en las traducciones, incluso si estructuras gramaticales similares están disponibles en los lenguajes de origen y destino. En el caso de inglés, también las diferencias entre sus variedades americano y británico pueden ser facilmente analizadas gracias al corpora.

Los cambios históricos del significado y gramática de las palabras son analizados como resultado del desarrollo del corpora. Y como la cantidad de viejos textos disponibles en formato electrónico es mucho más pequeña que la cantidad de textos contemporáneos el trabajo es relizable. Por lo tanto, las diferencias en los aspectos gramaticales concernientes a la voz pasiva fueron trazados y resulta que con el siglo 19 la voz pasiva en el lenguaje inglés comienza a ser utlizada más y más frecuentemente.

Cuando corpora escrito y hablado se hizo disponible, los linguistas comenzaron a analizarlo para verificar si había patrones o diferencias entre lo hablado y lo escrito. Parece que aparte de algunas características obvias como falsos comienzos y dudas que ocurren en lo hablado y no en lo escrito, la utilizacióon de un gran número de expresiones deícticas es más frecuente en los discursos orales. Probablemente esto es debido a los signos linguísticos extra de los que el lenguaje hablado es más vago. Adicionalmente ciertas características gramaticales manifestadas en lo hablado deben ser consideradas agramaticales en lo escrito.

Los linguistas utilizan una metodológía que intenta describir naturalmente las ocurrencias del lenguaje apoyadas en sus observacioens de gran cantidad de evidencia encontrada en el corpora. Más aún, las operaciones estadísticas están generalmente involucradas en el trabajo sobre el corpora, especialmente cuando las frecuencias de uso de algún aspecto léxico son medidas. Grandes bases de datos de ocurrencia natural del lenguaje han ayudado a realizar progresos en

el estudio de fraseología, especialmente cuando se ha descubierto que ciertos significados de las palabras se correlacionan con las estructuras gramaticales en donde son utilizadas.

La linguistica de corpus ha encontrado aplicaciones en muchos campos como: análisis crítico del discurso, estilismos, linguística forense así como traducciones y enseñanza del lenguaje.

En las traducciones es útil utilizar corpora paralelo que permite una mejor elección de equivalencias y estructuras gramaticales que podrían reflejar el significado deseado. Estudios adicionales sobre corpora revelaron que los traductores no traducen palabra por palabra sino unidades más grandes (cláusulas o sentencias).

Los estudios de corpora probablemente han tenido una gran influencia en la enseñanza del lenguaje. Primero que nada, han influído en la forma en que se hacen los diccionarios. Segundo los aprendices del lenguaje han sido estudiados para mejorar el conocimiento de los maestros. Y hoy en día los aprendices son alentados a hacer uso de corpora por sí mismos para incrementar su conocimiento. Inclusive la información de los resultados reunidos de corpora han influenciado el diseño y contenido de los libros de aprendizaje.

Los linguistas de corpus creen que un análisis confiable del lenguaje ocurre mejor en ejemplos recolectados de campo; contextos naturales y con interferencia experimental mínima. Dentro del corpus linguistico existen visiones divergentes en torno al nivel de las anotaciones. Desde John Sinclair abogando anotaciones mínimas y permitiendo a los textos «hablar por ellos mismos» a otros como el equipo de Survey of English Usage (University College, London) abogando anotaciones como un camino hacia un riguroso entendimiento linguístico.

0.2. Un poco de historia

El punto de inflexión en corpus linguístico moderno fué la publicación de Henry Kucera y W. Nelson Francis: Computational Analysis of Present-Day American English en 1967. Un trabajo basado en el análisis del corpus Brown, una compilación cuidadosamente seleccionada de inglés americano actual totalizando alrededor de 1 millón de palabras obtenidas de una amplia variedad de fuentes. Kucera y Francis sometieron este corpus a una gran variedad de análisis computacional desde el cual compilaron una rico y nutrido corpus combinando elementos de linguística, enseñanza de lenguaje, psicología, estadística y sociología. Una publicación clave adicional fué «Towards a description of English Usage» de Randolph Quirk (1960) en la que introdujo Survey of English Usage.

Poco después el publicador de Boston Houghton-Mifflin se acercó a Kucera para suministrarle la citación base de 1 millón de palabras para su nuevo diccionario American Heritage Dictionary (AHD), el primer diccionario que fué compilado utilizando corpus linguístico. EL AHD dió el paso innovador de combinar elementos prescriptivos (como debe utilizarse el lenguaje) con información descriptiva (como se usa actualmente).

Otros publicadores siguieron el ejemplo. El publicador inglés Collins creó y compiló el diccionario Cobuild utilizando el corpus Bank of English. Fué diseñado para usuarios que están aprendiendo inglés como lengua extranjera.

El corpus Brown también dió lugar a un número de corpora similarmente estructurada: el corpus LOB (1960, inglés británico), Kolhapur (inglés indio),

Wellington (inglés de Nueva Zelanda), Australian Corpus of English (inglés australiano) y el Flob corpus (1990, inglés británico).

Otro corpora representa muchos lenguajes, variedades y modos e incluye el International Corpus of English, el British National Corpus: una colección de 100 millones de palabras provenientes de textos escritos e inglés hablado, creado en los 1990s por un consorcio de publicadores, universidades (Oxford y Lancaster) y la British Library. Para inglés americano contemporáneo, el trabajo se ha centrado en el American National Corpus (más de 400 millones de palabras de inglés americano contemporáneo).

El primer corpus computarizado de lenguaje hablado transcripto fué construído en 1971 por el Montreal French Project, conteniendo 1 millón de palabras que inspiró a Shana Poplack a crear un corpus mucho más grande de Francés hablado.

Al lado de estos corpora de lenguaje vivo se encuentra corpora computarizado que también fué construído a partir de colecciones de textos en lenguajes antiguos. Como ejemplo tenemos la base de datos Andersen-Forbes de la biblia hebrea, desarrollada desde los años 1970, en donde cada cláusula es parseada utilizando grados que representan más de 7 niveles de sintaxis y cada segmento es etiquetado con 7 campos de información. El Quaric Arabic Corpus es un corpus anotado para el lenguaje árabe clásico del corán. Este es un proyecto reciente con múltiples capas de anotación incluyendo segmentación morfológica, etiquetado gramatical y análisis sintáctico utilizando dependencia gramatical.

0.3. Métodos

Corpus linguísticos han generado una cantidad de métodos de investigación intentando trazar un camino desde los datos hacia la teoría. Wallib y Nelson (2001) introdujeron lo que ellos llamaron la perspectiva 3A: anotación, abstracción y análisis.

La anotación consiste en la aplicación de un esquema a los textos. Las anotaciones incluyen marcado estructural, etiquetado gramatical, parsing y varias representaciones más. La abstracción consiste en la traducción (mapeo) de términos del esquema a términos en el modelo teórico. Típicamente incluye búsqueda linguística directa y tambi;en puede incluir aprendizaje por reglas para parsers. El análisis consiste de exploración estadística, manipulación y generalización desde los datos. El análisis también podría incluir evaluaciones estadísticas, optimización basada en reglas o métodos de descubrimiento del conocimiento.

La mayoría del corpora léxcio de hoy en día está anotado gramaticalmente. Sin embargo, incluso los corpus linguísticos que trabajan con texto plano no anotado inevitablemente aplican algún método para aislar términos que pueden ser interesantes en las palabras circundantes. En tales circunstancias anotación y abstracción son cominados en una búsqueda léxica.