|  |
| --- |
| Descubrimiento de información en textos |
| Tarea del tema 2: Estudio de distintos corpus |
|  |
| Gabriel Vázquez Torres |
|  |

Descubrimiento de información en textos

Tarea del tema 2: Estudio de distintos corpus

Contenido

[Descripción de la práctica 2](#_Toc501474743)

[Descripción de los corpus 3](#_Toc501474744)

[Brown Corpus 3](#_Toc501474745)

[Sussane Corpus 4](#_Toc501474746)

# Descripción de la práctica

En esta práctica vamos a realizar un análisis y estudio de diferentes corpus. Estos son los de BROWN, SUSANNE Y PENN TREEBANK.

En este documento se hablará sobre diferentes apartados:

1. Descripción de los corpus
2. Comparativa concisa de distintos aspectos que se consideren relevantes entre ellos:
   1. Tipo de etiquetado: etiquetado léxico (POS atgging), sintático, etc.
   2. Tamaño del corpus.
   3. Tamaño del conjunto de etiquetas.
   4. Temáticas incluidas.
   5. Procedencia de los textos: periódicos, transcripciones de habla, etc.
3. Breve análisis de cinco líneas como máximo sobre qué Corpus, si el de Susanne o el de Brown, es más apropiado en función de sus características para extraer información estadística significativa referente a cuales son las etiquetas y las parejas de etiquetas consecutivas que aparecen más frecuentemente en los textos.

# Descripción de los corpus

A continuación se hará una descripción de los corpus previamente comentados:

## Brown Corpus

La historia de la lingüística de corpus es muy corta. El punto de arranque obligado es la aparición en 1964 del “Brown University Estándar Corpus of Present-Day American English (**Brown Corpus**), de Francis y Kučera, que es el primer corpus concebido y construido para residir en una computadora y ser explotado mediante programación informática.

Tiene aproximadamente medio siglo, una historia corta y difícil en sus primeros tiempos y que coinciden casi exactamente con el primer período de expansión de la lingüística generativa, que en aquellos años mantenía una oposición radical a todo lo relacionado con hechos lingüísticos concretos, frecuencias, estadística, variación, etc. (Rojo)

Francis y Kučera publican en 1967 su obra “Computational Analysis of Present-Day American English”, que proporciona estadísticas de lo que hoy se conoce simplemente como “**Brown Corpus**”. El Brown Corpus es una selección cuidadosamente compilada del actual inglés americano con un total aproximado de un millón de palabras extraídas de una amplia variedad de fuentes. (Kucera).

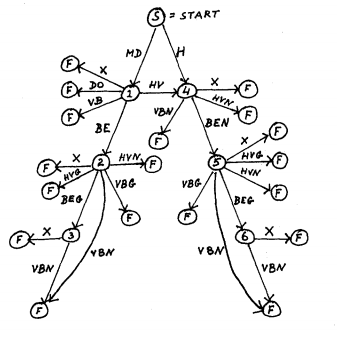


Ilustración 1: Representación en diagrama del corpus. Cada línea está etiquetada con un símbolo apropiado para la clase del nodo.

Brown Corpus ha sido la base de otros corpus posteriores a él como el Lancaster‐Oslo‐Bergen Corpus o SUSANNE. En un principio contenía 1.014.312 palabras y 15 categorías de texto. Es un corpus de tipo "POS Tagging" (Part‐Of‐Speech tagging, de análisis léxico) y posee 82 etiquetas distintas. (Lindebjerg, s.f.)

Actualmente sirve como conjunto de entrenamiento y de test en diferentes campos de la ciencia, e.g. PLN o biomedicina. La herramienta NLTK lo incorpora. (KHOKOVSKAIA)

En conclusión podríamos decir que Brown Corpus ha servido de inspiración para muchos investigadores. Fue la primera representación electrónica de un lenguaje y actualmente sigue siendo utilizado en diferentes ámbitos de la ciencia.

## Sussane Corpus