|  |
| --- |
| 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. [1]

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. [2].

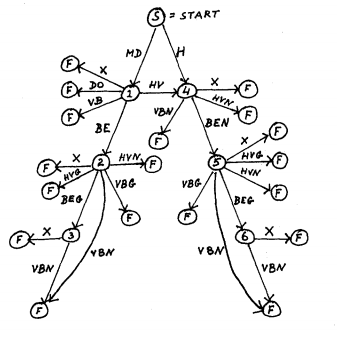


Ilustración 1: Representación en diagrama del corpus. Cada línea está etiquetada con un símbolo apropiado para la clase del nodo.

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. [3]

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. [4]

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

Susanne es la abrevación de *Surface and Underlying Structural ANalysis of Natural English*. Es un corpus procedente del Brown Corpus, que originariamente procedía de 64 de las 500 muestras del Brown Corpus y su versión inicial está formado por unas 130.000 palabras.

Al igual que el Brown Corpus, es de tipo *"POS Tagging"* (Análisis Léxico), y está formado por **353 etiquetas**, y las temáticas que contiene son las A,G,J y N del Brown Corpus (Prensa, Bellas Letras, Aprendidas y Ficción: Aventura y Occidental).

El Corpus SUSANNE se creó, con el patrocinio del Comité Económico y Social Research Council (Reino Unido), como parte del proceso de elaboración de una taxonomía completa del lenguaje de la ingeniería orientada y el esquema de anotación de la gramática (lógica y de la superficie) de Inglés. El Corpus SUSANNE, a pesar de haberse creado con el patrocinio de TEI, no cumple la normativa.

El esquema analítico SUSANNE ha sido desarrollado sobre la base de las muestras de ambos Inglés británico y americano. Fue inicialmente orientado hacia la lengua escrita solamente, y de hecho contiene muestras exclusivamente del lenguaje escrito. Sin embargo, en trabajos posteriores patrocinado por primera vez por el *Royal Signals and Radar Establishment*,  se produjeron extensiones al sistema para anotar los fenómenos distintivos estructurales del lenguaje hablado, y ha aplicado a estas muestras de los últimos Inglés hablado espontáneo (modificación mostrada en el Corpus CHRISTINE). La primera etapa del Corpus CHRISTINE, que incluye análisis de una equilibrada sección del Inglés hablado en todas partes del Reino Unido en la última década, fue lanzado en agosto de 1999 y es uno de los corpus orales para poder analizar el lenguaje hablado.

El Corpus SUSANNE abarca un subconjunto de aproximadamente 130.000 palabras del Brown Corpus de Inglés Americano, anotado, de acuerdo con el esquema de Susanne. Los motivos originales para la producción de esta base de datos incluye el de proporcionar mejores estadísticas para el análisis probabilístico, pero en este sentido, el Proyecto SUSANNE fue alcanzado después de su creación por los proyectos (en particular, Mitchell Marcus Pennsylvania proyecto Treebank) que han utilizado métodos cuasi-industrial para generar cuerpos mucho más grandes de material a analizar gramaticalmente. Sin embargo, el Corpus Susanne sí que mejora notablemente el análisis probabilístico en comparación con el Brown Corpus, aunque de esto hablaremos en el último punto del trabajo.

## Penn Treebank

El Treebank Penn, es un corpus de más de 4,5 millones de palabras de Inglés Americano. Durante la primera fase de tres años del Proyecto Penn Treebank (1989 - 1992), este corpus posee 2 tipos de etiquetado de, de tipo léxico y de tipo sintáctico. Sus orígenes están en la Universidad de Pennsylvania.

El conjunto de muestras del que está compuesto, procede de diferentes corpus distintos, concretamente de los siguientes: