

# Descubrimiento de Información en Textos

## Tarea 2

**José Alberto Benítez Andrades**

**71454586A**

**Descubrimiento de Información en Textos**

**Máster en Lenguajes y Sistemas Informáticos - Tecnologías del Lenguaje en la Web**

**UNED**

**7 de enero de 2011**

## 0. Introducción

La tarea del tema 2 para la asignatura de "*Descubrimiento de información en textos*" del *Máster en Lenguajes y Sistemas Informáticos, especialidad de Tecnologías del Lenguaje en la Web*, consiste en realizar una investigación y un estudio sobre tres corpus distintos: **BROWN, SUSANNE y PENN TREEBANK**. A continuación, en el documento actual, contestaré a los siguientes enunciados:

- Descripción de los corpus
- Comparativa concisa (tabla o similar) de distintos aspectos que consideres relevantes:
  - Tipo de etiquetado: etiquetado léxico (POS tagging), sintáctico, etc.
  - Tamaño del corpus,
  - Tamaño del conjunto de etiquetas,
  - Temáticas incluidas,
  - Procedencia de los textos: periódicos, transcripciones de habla, etc.
- Breve análisis (5 líneas como máximo) de qué corpus, 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.

## 1. Descripción de los distintos corpus: Brown, Susanne y Penn Treebank.

### Brown Corpus

En 1961/1963, Kucera y Francis publicó su obra clásica *Computational Analysis of Present-Day American English* (1967), que proporcionan las estadísticas básicas en lo que hoy se conoce simplemente como el *Brown Corpus*. El Brown Corpus fue una selección cuidadosamente compilada del actual Inglés Americano, por un total de alrededor de un millón de palabras extraídas de una amplia variedad de fuentes. Kucera y Francis realizaron una variedad de análisis computacionales, gracias a los cuales consiguieron compilar una obra rica y variada, que combina elementos de la lingüística, la psicología, la estadística y la sociología. Ha sido muy ampliamente utilizado en la lingüística computacional, y fue durante muchos años uno de los recursos más citados en el campo.

Poco después de la publicación del primer análisis léxico estadístico, la editorial Bostón Houghton-Mifflin Kucera tuvo una oferta millonaria para crear un nuevo diccionario *American Heritage Dictionary*.

El Brown Corpus ha sido la base de otros corpus posteriores a él como el Lancaster-Oslo-Bergen Corpus o el SUSANNE. El corpus original (1961) contenía 1.014.312 palabras muestra de 15 categorías de texto:

- A. DE PRENSA: Reportaje (44 textos)
- B. PRENSA: Editorial (27 textos)
- C. DE PRENSA: Comentarios (17 textos)
- D. RELIGIÓN (17 textos)
- E. HABILIDAD Y HOBBIES (36 textos)
- F. POPULAR LORE (48 textos)
- G. Bellas Letras (75 textos)
- H. VARIOS: Gobierno de los EE.UU. y los órganos de Lujo (30 textos)
- J. APRENDIDAS (80 textos)
- K. FICCIÓN: General (29 textos)
- L. FICCIÓN: Misterio y ficción de detectives (24 textos)
- M. FICCIÓN: Ciencia (6 textos)
- N. FICCIÓN: Aventura y Occidental (29 textos)
- P. FICCIÓN: Romance y Love Story (29 textos)
- R. HUMOR (9 textos)

El Brown Corpus es un corpus de tipo "POS Tagging" (Part-Of-Speech tagging, de análisis léxico) y posee **82 etiquetas distintas**.

### SUSANNE Corpus

Susanne es la abreviación 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:

- Dept. of Energy abstract
- Dow Jones Newswire stories
- Dept. of Agriculture bulletins
- Library of America texts
- MUC-3 messages
- IBM Manual sentences
- WBUR radio transcripts
- ATIS sentences
- Brown Corpus, retagged

Tiene 36 etiquetas de análisis léxico, además de 12 etiquetas para puntuaciones y símbolos. Y 14 etiquetas de tipo sintáctico además de 4 elementos nulos. De los 3 Corpus que analizamos en este trabajo, este es el más nuevo y el mejor en el sentido del análisis probabilístico del lenguaje, ya que es más sencillo de analizar que los dos anteriores.

## 2. Comparativa de distintos aspectos relevantes.

### Tipo de etiquetado

Corpus	Etiquetado
Brown	Etiquetado léxico
Susanne	Etiquetado léxico
Penn Treebank	Etiquetado léxico y sintáctico.

En el corpus **Brown** y en el corpus **Susanne**, el tipo de etiquetado que tienen es de POS Tagging (Part Of Speech Tagging) o lo que es lo mismo, etiquetado léxico. Sin embargo el corpus **Penn Treebank** posee un etiquetado mixto, de tipo léxico y de tipo sintáctico.

### Tamaño del corpus

Corpus	Tamaño del corpus
Brown	500 muestras de 2.000 o más palabras (1.014.312 palabras en total)
Susanne	64 Muestras de 2.000 o + palabras cada una (130.000 palabras)
Penn Treebank	4.885.798 palabras en total.

El corpus **Susanne** posee 64 muestras obtenidas de las muestras del corpus **Brown**, y el corpus **Penn Treebank**, obtuvo esas palabras de distintos textos anteriormente numerados.

### Tamaño del conjunto de etiquetas

Corpus	Tamaño del conjunto de etiquetas
Brown	82 divididas en 6 partes A. Partes de la oración Nombre, común y propio, verbo, adjetivo.... B. Función de las palabras: determinantes, preposiciones, conjunciones... C. Palabras individuales importantes: no, infinito existencial, la forma del verbo. D. Las marcas de puntuación de importancia sintáctica. E. Morfemas flexivos. F. Dos etiquetas (FM y NC) PALABRAS extranjera o citada.
Susanne	353 wordtags (sin contar las etiquetas para expresiones gramaticales)
Penn Treebank	36 y 12 para puntuaciones y símbolos en el etiquetado léxico. Para el etiquetado sintáctico 14 y además 4 más para elementos nulos.

## Temáticas incluidas

Corpus	Temáticas incluidas
Brown	A. DE PRENSA: Reportaje (44 textos) B. PRENSA: Editorial (27 textos) C. DE PRENSA: Comentarios (17 textos) D. RELIGIÓN (17 textos) E. HABILIDAD Y HOBBIES (36 textos) F. POPULAR LORE (48 textos) G. Bellas Letras (75 textos) H. VARIOS: Gobierno de los EE.UU. y los órganos de Lujo (30 textos) J. APRENDIDAS (80 textos) K. FICCIÓN: General (29 textos) L. FICCIÓN: Misterio y ficción de detectives (24 textos) M. FICCIÓN: Ciencia (6 textos) N. FICCIÓN: Aventura y Occidental (29 textos) P. FICCIÓN: Romance y Love Story (29 textos) R. HUMOR (9 textos)
Susanne	Las temáticas que contiene son obtenidas del corpus Brown: A. De Prensa. G. Bellas Letras. J. Aprendidas N. Ficción : Aventura y Occidental.
Penn Treebank	36 y 12 para puntuaciones y símbolos en el etiquetado léxico. Para el etiquetado sintáctico 14 y además 4 más para elementos nulos.

## Procedencia

Corpus	Temáticas incluidas
Brown	R. HUMOR (9 textos)
Susanne	El Susanne procede del corpus Brown, de 64 de las 500 muestras que posee el corpus Brown, y fue creado para mejorar su análisis probabilístico.
Penn Treebank	Procede de los siguientes documentos: <ul style="list-style-type: none"> <li>▪ Dept. of Energy abstract</li> <li>▪ Dow Jones Newswire stories</li> <li>▪ Dept. of Agriculture bulletins</li> <li>▪ Library of America texts</li> <li>▪ MUC-3 messages</li> <li>▪ IBM Manual sentences</li> <li>▪ WBUR radio transcripts</li> <li>▪ ATIS sentences</li> <li>▪ Brown Corpus, retagged</li> </ul>

**3. Breve análisis (5 líneas como máximo) de qué corpus, el de Susanne o el de Brown, es más apropiado en función de sus características para extraer información estadística significativa referente a cuáles son las etiquetas y las parejas de etiquetas consecutivas que aparecen más frecuentemente en los textos.**

Es mejor el SUSANNE que el BROWN CORPUS. *Susanne* posee un conjunto de etiquetas mucho más preciso, mucho más granular y más fácil de interpretar por las personas que las etiquetas del corpus *Brown*. Por ejemplo la etiqueta CSN que se utiliza en *Susanne* es equivalente a utilizar la etiqueta *cs* y la preposición *in* en el corpus *Brown*. En conclusión, son más sencillas de interpretar las etiquetas de *Susanne* que las de *Brown*.