Parte I - Corpus y sus características

Descripción de los corpus Brown, Susanne y Penn Treebank

Brown Corpus

El Brown Corpus es el primer gran corpus de inglés moderno (compilado en 1961 en la Universidad de Brown, EE.UU), compuesto por 1 millón de palabras tomadas de 500 textos de diferentes géneros (prensa, ficción, ensayo, religión, técnica, etc.). Cada fragmento de texto es de aproximadamente 2.000 palabras cada una, que en total suman alrededor de 1.014.000 palabras.

Cada palabra está etiquetada con su categoría gramatical o POS (Part-of-Speech), usando un conjunto de 80 etiquetas. Es representativo del inglés escrito de la época y ha sido fundamental para estudios de frecuencia y análisis léxico.

Susanne Corpus

El Susanne Corpus es un subconjunto detalladamente anotado del Brown (unas 128.000 palabras), desarrollado en los años 90. Aporta una anotación mucho más completa, incluyendo POS y análisis sintáctico estructural (árboles de constituyentes), con un esquema de 353 etiquetas que diferencia matices morfosintácticos. Está pensado para estudios gramaticales profundos y es utilizado sobre todo para evaluar parsers y análisis sintáctico.

Penn Treebank

El Penn Treebank es un corpus de finales de los 80 y principios de los 90, compuesto principalmente por 1 millón de palabras de artículos del Wall Street Journal (noticias económicas y generales). Incluye etiquetado POS (36 etiquetas básicas, estándar Penn). Este esquema de etiquetado morfosintáctico es ligeramente distinto al de Brown: consta de 36 etiquetas POS básicas (sustantivo singular = NN, plural = NNS, verbo base = VB, verbo pasado = VBD, determinante = DT, etc.), a las que se suman etiquetas para signos de puntuación y símbolos especiales (hasta un total de unos 45 símbolos si incluimos puntuación). Este corpus también posee anotación sintáctica en árboles de constituyentes. Es el corpus de referencia para desarrollo y evaluación de parsers estadísticos y modelos de etiquetado POS en inglés contemporáneo. No es de acceso libre.

Aspecto	Brown Corpus (1961)	Susanne Corpus (1992)	Penn Treebank (1995*)
Tipo de etiquetado	Etiquetado léxico (categorías gramaticales POS).	Etiquetado léxico + sintáctico completo	Etiquetado léxico + sintáctico (árboles de constituyentes, esquema Penn).
Tamaño del corpus	~1.000.000 de palabras (500 textos de ~2000 c/u).	~128.000 de palabras (submuestra del Brown).	~1.000.000 de palabras
Tamaño del conjunto de etiquetas	~80 etiquetas POS (Brown tagset).	353 etiquetas distintas (esquema Susanne para POS y funciones).	36 etiquetas POS básicas (estándar Penn; ~45 incl. puntuación) + categorías sintácticas.
Temáticas incluidas	Muy variadas: prensa, textos académicos, técnicos, ficción, religión, ensayos, humor, etc.	Variedad similar a Brown (muestra representativa de los mismos géneros del Brown, aunque de menor tamaño).	Menos diversa: principalmente artículos periodísticos financieros y de noticias generales (Wall Street Journal).
Procedencia de los textos	Textos escritos publicados en EE.UU. en 1961.	Textos escritos de 1961 (EE.UU.). *Subset* del Brown Corpus original.	Textos periódísticos (Wall Street Journal, EE.UU., 1987-89) mayoritariamente

Análisis comparativo: Brown vs. Susanne

Para extraer información estadística sobre frecuencias de etiquetas léxicas y parejas de etiquetas consecutivas, el Brown Corpus es más apropiado que Susanne porque:

- Tamaño mayor: permite obtener estadísticas más fiables y representativas.
- Variedad temática: abarca más géneros y registros, reflejando el uso general del inglés escrito.

- Etiquetado POS menos granular: facilita el análisis y agrupamiento de etiquetas léxicas y sus combinaciones más frecuentes, sin perder información relevante para el estudio léxico.

Susanne es más útil para estudios gramaticales complejos y análisis sintáctico detallado, pero para análisis de frecuencias de etiquetas y bigramas, Brown ofrece una base más robusta y generalizable.

Parte II - Anotaciones de documentos

En la realización de este ejercicio, se ha utilizado el IDE Visual Studio Code.

Para poder determinar si el archivo .xml está bien formado, se utiliza el siguiente comando de `xmllint` (herramienta de línea de comandos para analizar y manipular archivos XML):

```
xmllint --noout Practica_Tema3_ejemplo1.xml
```

El documento XML original no estaba bien formado debido a los siguientes errores:

- Error en línea 45: La etiqueta <note> no estaba cerrada correctamente. Faltaba el cierre </note> antes de abrir la siguiente etiqueta <note>. En XML, toda etiqueta que se abre debe cerrarse correctamente. La estructura correcta requiere que cada <note> tenga su correspondiente </note>.
- Error en línea 96: Se usó la etiqueta incorrecta <pbb> en lugar de <pb>. Esto es un error tipográfico. Según el estándar TEI (Text Encoding Initiative), la etiqueta correcta para marcar un salto de página es <pb> (page break), no <pbb>.
- Error en línea 114: El atributo estaba mal escrito como reff="char:EOLhyphen" en lugar de ref="char:EOLhyphen". Esto es un error tipográfico en el nombre del atributo. El atributo estándar para referencias en XML/TEI es ref, no reff.
- Error en línea 135: La etiqueta no estaba cerrada correctamente. Faltaba el cierre al final del párrafo. Toda etiqueta de párrafo debe cerrarse con para mantener la estructura jerárquica del XML.

A la hora de comprobar si dicho documento XML es un documento válido o conforme con la DTD, se utiliza el siguiente comando (primero es necesario referenciar el DTD de manera interna usando `<!DOCTYPE TEI SYSTEM "tei_all.dtd">`):

```
xmllint --valid --noout Practica_Tema3_ejemplo1.xml
```

El documento XML original no era válido conforme al DTD debido a:

- Atributos de namespace no declarados: Los atributos xmlns:tei="http://www.tei-c.org/ns/1.0" y xmlns="http://www.tei-c.org/ns/1.0" no están declarados en la DTD tei_all.dtd. Se eliminan estos atributos ya que las DTDs clásicas no soportan namespaces XML de la misma manera que los esquemas XML Schema
- Estructura incorrecta en publicationStmt: El elemento publicationStmt dentro de biblFull solo contenía <pubPlace> y <date>, pero la DTD requiere al menos un elemento publisher, distributor, o authority antes de los otros elementos. Se agrega <publisher>Unknown</publisher> para cumplir con el modelo de contenido requerido

- Atributo num no declarado: El elemento editorialDecl tenía un atributo num="4" que no está declarado en la DTD. Se elimina el atributo num del elemento editorialDecl