Estudio de corpus y anotaciones XML

Asignatura: Descubrimiento de información en textos.
Profesores: M. Lourdes Araujo Serna, Raquel Martinez, Unanue
Alumno: Macià Salvà Salvà

Indice

Parte I: Corpus y sus características

Corpus Brown

El corpus Brown alberga más de un millón de palabras, lo que lo hace una herramienta poderosa para llevar a cabo investigaciones detalladas sobre el lenguaje escrito en inglés. Además, cada una de estas palabras se registra con categorías léxicas, lo que facilita la realización de análisis no únicamente sobre las palabras en sí, sino también acerca de su papel en la estructura gramatical de una frase. Estos, proporcionan datos exhaustivos acerca del papel que juega cada palabra, favoreciendo la investigación en áreas como la lingüística computacional, la instrucción del idioma y la creación de aplicaciones para el procesamiento del lenguaje natural.

Una de las particularidades de este corpus es su organización en archivos que abarcan 15 temas distintos. Esta variedad temática proporciona al corpus de Brown una representación más extensa, dado que incluye estilos variados, facilitando la realización de comparaciones y análisis en contextos particulares, o bien, analizar la aplicación del lenguaje en distintos campos de comunicación.

Para un análisis más exacto, el corpus emplea un sistema de clasificación que abarca un total de 87 categorías gramaticales. Dentro de estas categorías se incluyen:

- **AT**: que se refiere a los artículos.
- **BE**: utilizada para marcar el verbo "to be" cuando se encuentra en su forma infinitiva o en modo imperativo.
- **CC**: que designa las conjunciones copulativas.
- **DT**: etiqueta que abarca tanto a los determinantes como a los pronombres.

Estas etiquetas permiten un etiquetado detallado de cada palabra, facilitando el análisis gramatical y sintáctico de los textos. Este nivel de precisión es crucial para aplicaciones como el etiquetado automático, la creación de modelos lingüísticos y la enseñanza de la gramática. De esta manera, el corpus de Brown se convierte en una herramienta para aplicaciones prácticas en el campo del procesamiento del lenguaje natural.

```
The/AT Fulton/NP County/NP Grand/NP Jury/NP said/VBD Friday/NR an/AT
investigation/NN of/IN Atlanta/NP 's/\$ recent/JJ primary/NN election/NN
produced/VBD "/" no/AT evidence/NN ''/'' that/CS any/DTI irregularities/NNS
took/VBD place/NN ./. The/AT jury/NN further/RBR said/VBD in/IN term-end/NN
presentments/NNS that/CS the/AT City/NP Executive/NP Committee/NP ,/, which/WDT
```

```
had/HVD over-all/JJ charge/NN of/IN the/AT election/NN ,/, "/" deserves/VBZ the/AT
  praise/NN and/CC thanks/NNS of/IN the/AT City/NP of/IN Atlanta/NP ''/'' for/IN
  the/AT manner/NN in/IN which/WDT the/AT election/NN was/BEDZ conducted/VBN ./.
```

Imagen 1. Corpus Brown.

Corpus Susanne

El corpus está compuesto por 130,000 palabras, extraídas de un subconjunto de 64 archivos del corpus de Brown. Cada palabra está anotada con categorías léxicas y análisis sintáctico detallado.

Para cada palabra del texto, se incluye una línea de anotaciones que proporciona la siguiente información:

- **Referencia**: indica el nombre del archivo, la línea y la posición dentro de la línea.
- **Estado**: señala si la palabra es una abreviatura o símbolo.
- **Categoría gramatical**: clasifica la palabra según su función en la oración.
- **Palabra**: el término tal como aparece en el texto.
- **Lema**: la forma básica o diccionario de la palabra.
- **Análisis sintáctico**: describe su papel dentro de la estructura oracional.

Este formato estructurado permite un análisis exhaustivo de cada elemento del texto, facilitando tanto estudios lingüísticos como aplicaciones prácticas en el procesamiento de lenguaje natural.

Un ejemplo es el siguiente.

Referencia	Estado	Categoría gramatical	Palabra	Lema	Análisis sintáctico
N06:0180.12	-	NN1u	Baldness	baldness	[S[Ns:s.Ns:s]
N06:0180.15	-	VBDZ	was	be	[Vsu
N06:0180.18	-	VVGt	attacking	attack	.Vsu]
N06:0180.21	-	APPGm	his	his	[Ns:o.
N06:0180.24	-	NN1c	pate	pate	.Ns:o]S]

Corpus Penn Treebank

Este corpus contiene una gran cantidad de textos anotados con información gramatical y sintáctica, extraídos principalmente de fuentes como el Wall Street Journal. El corpus incluye análisis detallados de estructura sintáctica (en forma de árboles sintácticos) y etiquetado morfosintáctico (categorías gramaticales). Su objetivo es proporcionar una base para el entrenamiento y evaluación de modelos automáticos de lenguaje, y ha sido clave en avances en áreas como el análisis sintáctico, el reconocimiento de voz y la traducción automática.\

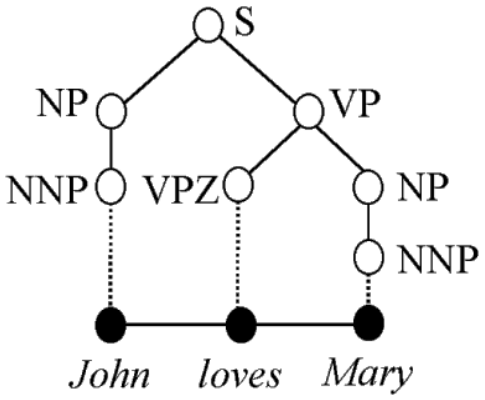

Imagen 3. Corpus Brown

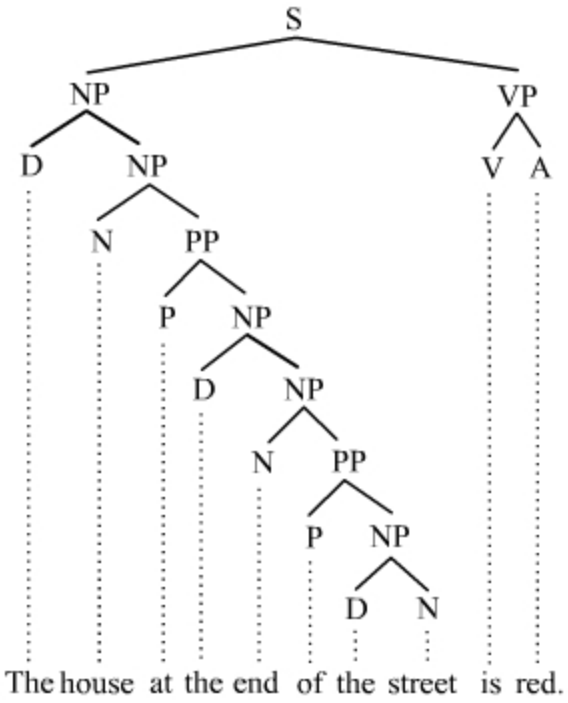

Imagen 4. Corpus Brown

Comparativas

	Brown	Susanne	Penn Treebank
Etiquetado	Después de cada palabra se pone una barra ‘/’ con la etiqueta correspondiente.	Anotado con categorías léxicas y análisis sintáctico.	Etiquetado en Árbol.
Tamaño del corpus	El tamaño del corpus resultante será prácticamente el mismo que el del corpus de entrada, ya que solo se añade una barra (/) y su correspondiente clase después de cada palabra.	Este será el que ocupe más espacio, ya que almacena 6 registros por cada palabra.	Este tendrá un formato similar al de Brown, ya que organiza una estructura en árbol para cada palabra del texto.
Tamaño del conjunto de	87 etiquetas.	130.000 palabras	

etiquetas.			
Temáticas incluidas.	Ficción, ciencia, religión, periodismo, ensayos, etc.		Textos periodísticos
Procedencia de los textos.	Textos escritos en inglés americano de diversas fuentes, como ficción, no ficción, periódicos y revistas		Textos periodisticos de Wall Street Journal.

Análisis entre Susanne y Brown (estadística)

El corpus de Brown y el de Susanne son dos recursos lingüísticos que ofrecen diferentes características y enfoques para el análisis estadístico de etiquetas léxicas. El corpus de Brown, siendo un conjunto más antiguo y de mayor tamaño, proporciona una variedad de géneros y estilos textuales, lo que permite un análisis más amplio y representativo del lenguaje en diferentes contextos. Además, su diseño y estructura favorecen la identificación de patrones y la extracción de frecuencias de etiquetas léxicas.

Por otro lado, el corpus de Susanne se centra en el inglés contemporáneo y es más específico en su enfoque, lo que puede resultar útil para estudios más dirigidos hacia un tipo particular de lenguaje. Sin embargo, su tamaño y diversidad son menores que los del corpus de Brown.

En términos de extracción de información estadística significativa sobre etiquetas léxicas y parejas de etiquetas léxicas consecutivas, el corpus de Brown es más apropiado debido a su amplitud y diversidad, lo que favorece un análisis más robusto y generalizable de las frecuencias y patrones lingüísticos.

Parte II: Anotaciones de documentos

Parte 2.1

Comprobar si dicho documento XML es un documento bien formado o bien construido. Explicar porque lo es o no, haciendo las correcciones necesarias en caso de que no lo sea y justificándose.

El primer error en el formato XML es la etiqueta de cierre de "note", como se muestra en la *Imagen 5*.

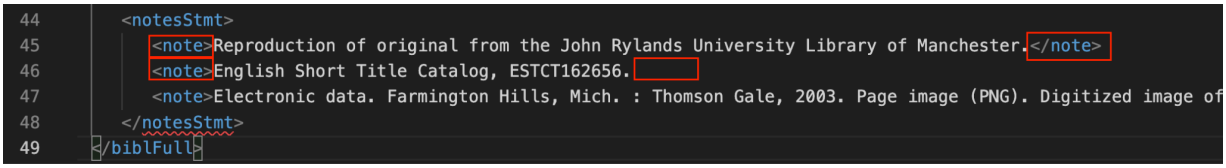

Imagen 5. Error en la línea 46, falta cerrar la etiqueta 'note'.

Este error se corrige simplemente añadiendo la etiqueta de cierre, como se muestra en la *Imagen 6*.

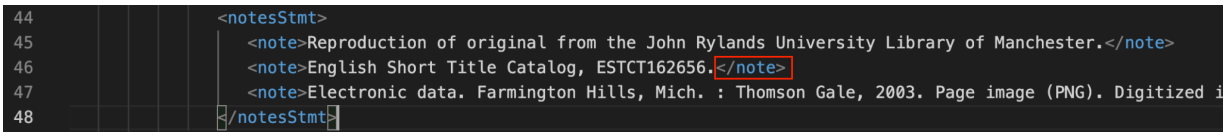

Imagen 6. Corrección de la etiqueta.

A continuación, se presenta otro error en la estructura del XML: falta nuevamente la etiqueta de cierre, ver *Imagen 7*.

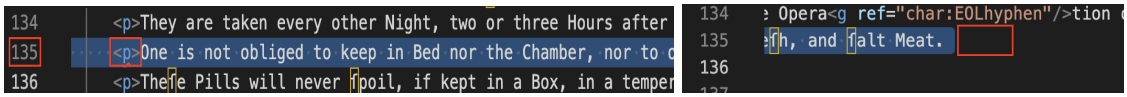

Imagen 7. Ausencia de cerrado de la etiqueta 'p'.

Este error se corrige añadiendo la etiqueta de cierre `</p>`.

Parte 2.2

Comprobar si dicho documento XML es un documento válido o conforme con la DTD de TEI tei all.dtd que se encuentra en el repositorio Explicar porque lo es o no en cada caso.

El primer error que encontramos en el formato DTD de nuestro XML lo podemos ver en la *Imagen 8*.

Imagen 8. Error del fichero xml.

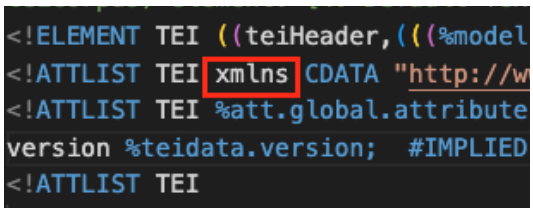

*Imagen 9. Definición del atributo **xlmns** en nuestro dtd.*

En la *Imagen 8*, se puede observar que dentro del elemento TEI se ha definido un atributo llamado `xmlns:tei`. Sin embargo, nuestro editor de código lo señala como un error, ya que la etiqueta `xmlns:tei` no existe. Por otro lado, en la *Imagen 9* se encuentra definida una etiqueta `xmlns`. Por lo tanto, procederemos a cambiar el nombre de `xmlns:tei` a `xmlns`.

Imagen 10. Solución al problema de la etiqueta xmlns.

El siguiente error que encontramos está en la estructura de la etiqueta `publicationStmt`.

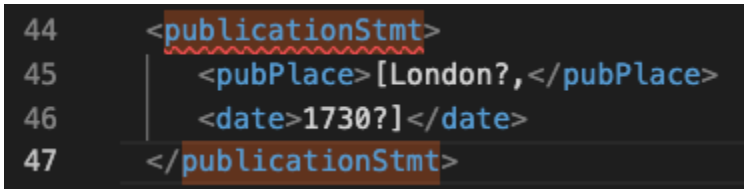

Imagen 11. Error en la etiqueta `publicationStmt`.

Si vamos a la definición de `publicationStmt` encontramos que

Imagen 12. Definición en la dtd de `publicationStmt`.

Esto significa que entre la etiqueta *publicationStmt* debe haber al menos un publisher, distributor, o authority al principio.

Como podemos observar en la *Imagen 10*, no hay ningún elemento correspondiente a publisher, distributor o authority. Además, al revisar el contenido dentro de las etiquetas pubPlace y date, parece que también son incorrectas o que no tienen mucho sentido. Por otro lado, al mirar una etiqueta un poco más arriba, la publicationStmt, se observa que coincide con los titulos y los autores, ver *Imagen 12* y *Imagen 13.*

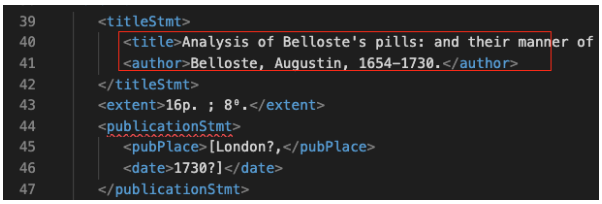

Imagen 13. Vemos arriba el título y el autor.

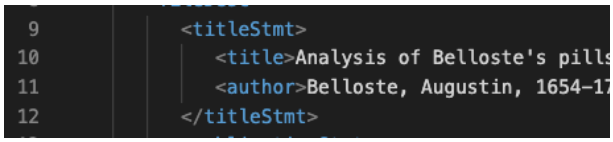

Imagen 14. Vemos al principio del fichero el mismo título y el mismo autor.

Todo esto me lleva a pensar que, dado que ya tenemos la etiqueta distributor definida más arriba, podemos integrarla en la etiqueta que presenta problemas, ya que es probable que la información contenida en su interior sea la correcta.

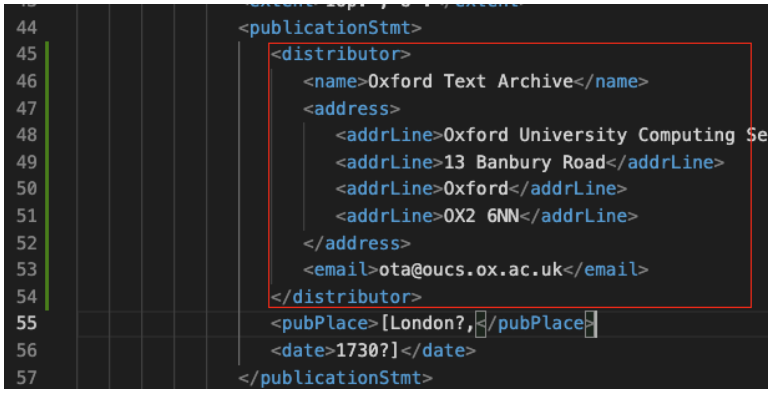

Imagen 15. Añadimos el mismo bloque distributor.

En el siguiente error, encontramos una inconsistencia relacionada con la declaración "num".

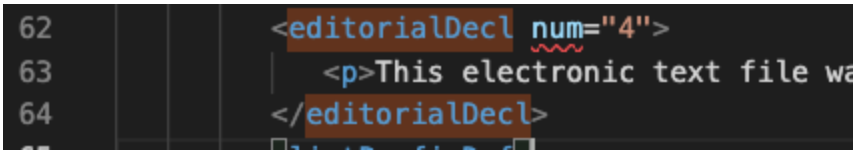

Imagen 16. Solución al problema de la etiqueta xmlns.

Si revisamos la definición de `editorialDecl`, podemos ver que no se declara ningún atributo llamado "num".

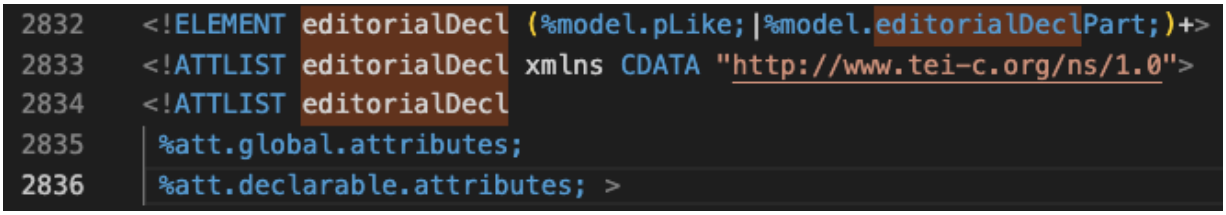

Imagen 17. Solución al problema de la etiqueta xmlns.

Entonces deducimos que ese num es una equivocación ya que no esta declarado en el fichero tdt, por tanto borramos la etiqueta num y queda como en la *Imagen 17.*

Imagen 17. Solución al problema del atributo num.

El siguiente error que encontramos es que nuestro `div` se resalta en color rojo, y dentro de este, hay un elemento que también está marcado en rojo.

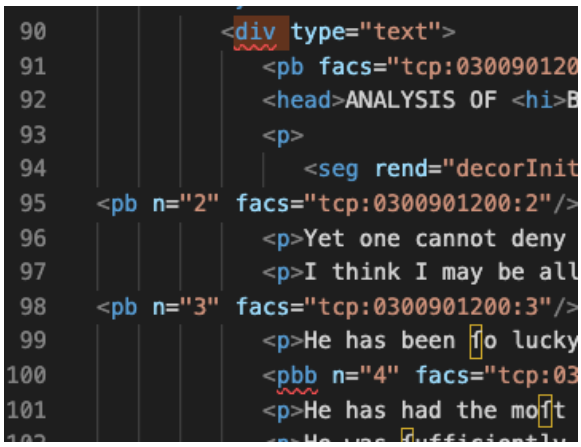

Imagen 18. Error en el etiquetado pbb.

Si buscamos en las definiciones de nuestra DTD, notamos que no tenemos ningún elemento definido como `pbb`.

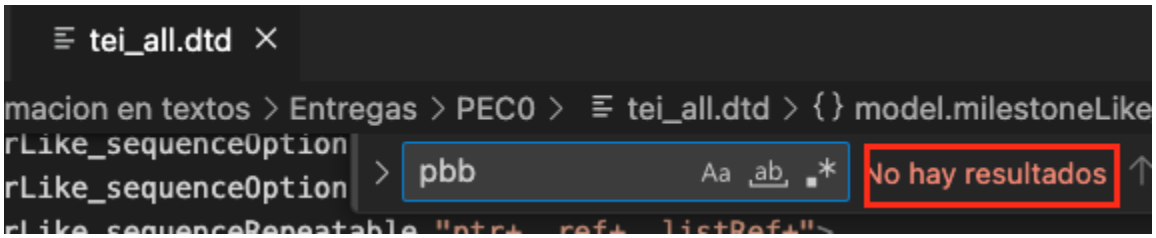

Imagen 19. Vemos que el elemento pbb no está definido en la DTD.

Sin embargo, más adelante encontramos que el elemento `pb` sí está definido en nuestra DTD. Por lo tanto, para solucionarlo, realizaremos el siguiente cambio, ver *Imagen 20*.

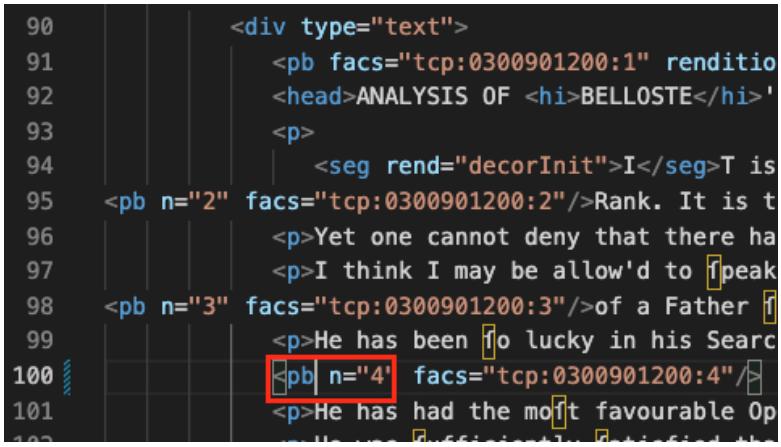

Imagen 20. Solucionamos el etiquetado de pbb a pb.

El último error que encuentro se muestra en la imagen 21.

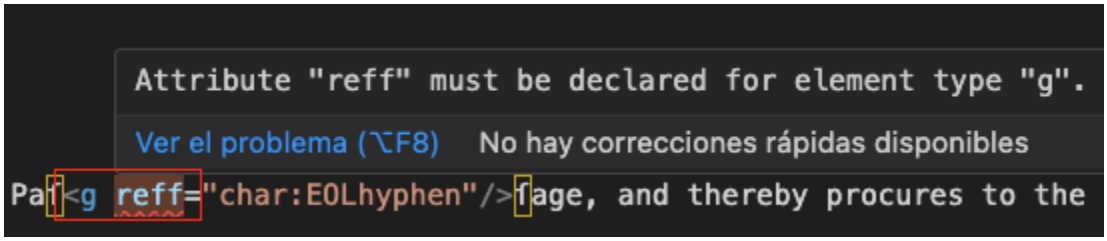

Imagen 21. Error en el etiquetado.

El problema se soluciona simplemente cambiando "reff" por "ref".

Imagen 22. Corrección del error.

Para realizar esta tarea, utilizo el editor de código Visual Studio Code junto con la extensión XML, que se muestra en la *Imagen 23*.

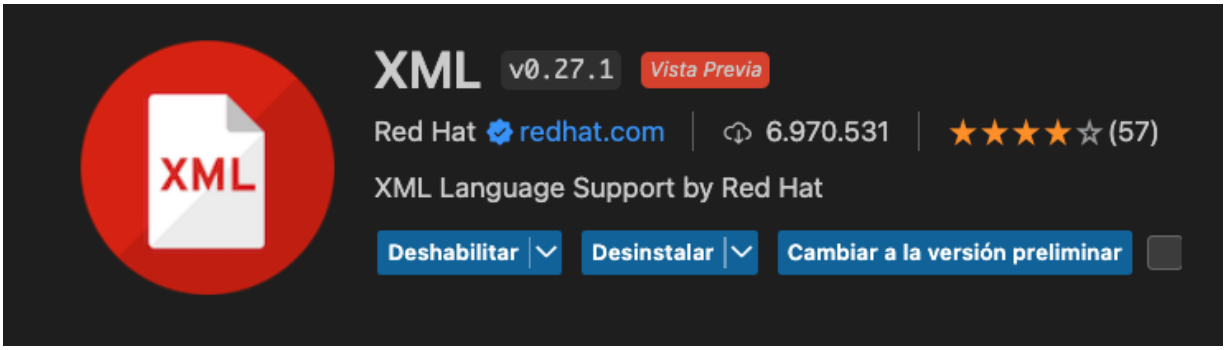

Imagen 23. Extension XML de VSCode.