|  |
| --- |
| Descubrimiento de información en textos |
| Tarea del tema 3: Estudio de estándares de anotaciones |
|  |
| Gabriel Vázquez Torres |
|  |

Descubrimiento de información en textos

Tarea del tema 3: Estudio de estándares de anotaciones

Contenido

[Descripción de la práctica 2](#_Toc502696683)

[Descripción de los corpus 3](#_Toc502696684)

[Brown Corpus 3](#_Toc502696685)

[Sussane Corpus 4](#_Toc502696686)

[Penn Treebank 5](#_Toc502696687)

[Comparativa de distintos aspectos 6](#_Toc502696688)

[Tipo de etiquetado 6](#_Toc502696689)

[Tamaño del corpus 6](#_Toc502696690)

[Tamaño del conjunto de etiquetas 6](#_Toc502696691)

[Temáticas incluidas 7](#_Toc502696692)

[Procedencia 8](#_Toc502696693)

[Breve análisis de qué corpus es el más apropiado diferenciando entre el corps de Brown y el de Sussane. 9](#_Toc502696694)

[Bibliografía 10](#_Toc502696695)

# Descripción de la práctica

A partir de un fichero en formato xml:

1. Comprobar si dicho documento XML es un documento conforme con la DTD de TEI Lite o con un subconjunto conforme con TEI Lite. Explicar por qué lo es o no.
2. En caso de que no lo sea, realizar las adiciones y sustituciones de elementos y atributos necesarias para que lo sea. Explica el porqué de cada adición y sustitución. Tened en cuenta que cuando un elemento o atributo no sea correcto para TEI Lite, habrá que intentar sustituirlo por otro que sí lo sea y que permita mantener la anotación.
3. Diseña una DTD que sólo contemple los elementos y atributos XML que hayas utilizado del TEI Lite.

# Apartado 1

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

# Apartado 2

# Apartado 3

# Bibliografía

|  |  |
| --- | --- |
| [1] | G. Rojo. [En línea]. Available: https://gramatica.usc.es/~grojo/Publicaciones/Lgca\_corpus\_lgca\_espanol.pdf. |
| [2] | F. y. Kucera, «aclweb,» [En línea]. Available: http://www.aclweb.org/anthology/C80-1006. |
| [3] | A. Lindebjerg. [En línea]. Available: http://www.hit.uib.no/icame/brown/bcm.html. |
| [4] | O. KHOKOVSKAIA. [En línea]. Available: http://radio.feld.cvut.cz/conf/poster/proceedings/Poster\_2017/Section\_HS/HS\_018\_Kholkovskaia.pdf. |
| [5] | [En línea]. Available: https://en.wikipedia.org/wiki/Brown\_Corpus. |
| [6] | [En línea]. Available: https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus\_ling/content/corpora/list/public/susanne.html. |
| [7] | [En línea]. Available: https://www.grsampson.net/SueDoc.html. |
| [8] | [En línea]. Available: https://research.csc.fi/-/susanne-corpus. |
| [9] | [En línea]. Available: http://www.lllf.uam.es/~sandoval/UAMTreebank.html. |
| [10] | [En línea]. Available: http://clic.ub.edu/. |
| [11] | [En línea]. Available: https://es.wikipedia.org/wiki/TreeBank. |
| [12] | «books.google,» [En línea]. Available: https://books.google.es/books?id=r3xyBgAAQBAJ&pg=PA238&lpg=PA238&dq=thematics+penn+treebank&source=bl&ots=rI5UVySG3q&sig=gVXpUTbH7dacg8NVA52idx\_LtHM&hl=en&sa=X&ved=0ahUKEwiPuaaB5rnYAhWKwbwKHWy1AqQQ6AEIMDAB#v=onepage&q=thematics%20penn%20treebank&f=false. |
| [13] | [En línea]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.8216&rep=rep1&type=pdf. |
| [14] | [En línea]. Available: https://www.grsampson.net/SueDoc.html. |
| [15] | [En línea]. Available: https://www.ling.upenn.edu/courses/Fall\_2003/ling001/penn\_treebank\_pos.html. |
| [16] | [En línea]. Available: https://www.clips.uantwerpen.be/pages/mbsp-tags. |
| [17] | [En línea]. Available: https://catalog.ldc.upenn.edu/docs/LDC95T7/cl93.html. |
| [18] | Ron Kohavi, George H. John, «Wrappers for feature subset selection,» Mountain View, 1995. |
| [19] | José Hernández Orallo, M. José Ramírez Quintana César, Introducción a la Minería de Datos. |