Breve manual de usuario de la herramienta QtNLP
4 de octubre de 2015
versión 1.0 de QtNLP
Autor: Abel Meneses Abad

El presente documento tiene como objetivo guiar a los especialistas de cualquier dominio del conocimiento para aprender a utilizar la herramienta de creación de corpus lingüísticos "*QtNLP v1.0".*

Índice de contenido

Resumen

QtNLP es una aplicación de escritorio del tipo Front-End para la construcción de corpus, construcción de diccionarios especializados y análisis de algoritmos de detección de similaridad. Está desarrollada en Qt y Python y es de uso libre para investigaciones.

TNLP es un corpus anotado de textos basado en el formato XML. Fue elaborado de forma genérica para la construcción de múltiples corpus lingüísticos, dirigidos a utilizarse como recurso en la solución de varios problemas del área del Procesamiento de los Lenguajes Naturales (NLP). Mayormente re-utilizables en el entrenamiento y prueba de algoritmos. Una de sus características fundamentales es la de contener los textos originales para la extracción de los objetos de estudio, similar al corpus PAN.[1] Este diseño puede ser usado en problemas que requieran una lista de casos conformados por dos partes o elementos, y anotaciones para cada caso que permitan verificar una propiedad entre estas dos partes.

Ejemplo corto: descripción de un caso

Problema NLP: similaridad semántica entre dos oraciones.

Oración 1: ... compilar el código en un pc no estándar, y obtener el ejecutable.

Oración 2: ... construir un binario en un procesador ejecutando funciones en ensamblador.

Propiedad: Tipo de Paráfrasis = semántica.

Tomado del README.html del corpus TNLP (Meneses-Abad, 2015)

Cargar Corpus

Pasos:

1. Copiar la carpeta de la versión de QtNLP nombrada “TNLP_branch_XETID” para su usuario.

2. Copiar los datos con los que trabajará
   a) sí trabajara en las temáticas Base de Datos – IGSW: copie la carpeta TNLP_group1
   b) sí trabajara en las temáticas de IA – Diseño-Test de Algoritmos: copie la carpeta TNLP_group2
   Nota: cada carpeta contiene una carpeta *susp* y una carpeta *src* + TNLP_base.xml

3. Coloque *susp* + *src* + *TNLP_base.xml* dentro de la ruta ~/TNLP_branch_XETID/data/corpuses/TNLP

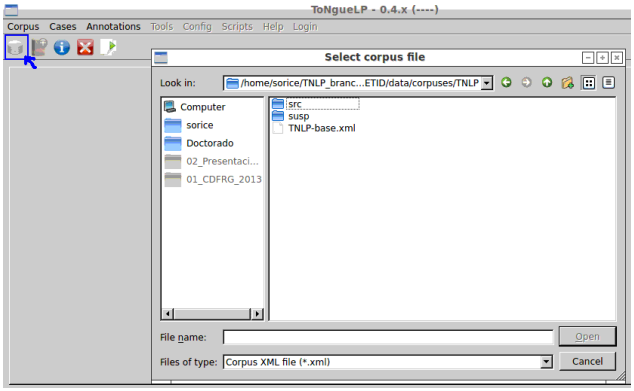
4. Ejecute en la consola parado en la ruta ~/TNLP_branch_XETID la siguiente orden:

```
~TNLP_branch_XETID$ python ToNgueLP.py
```

5. Al aparecer la aplicación apriete el botón  *ó click en “Corpus Menú” → “Load Corpus”.*

Y cargue el xml copiado en la carpeta ~data/...

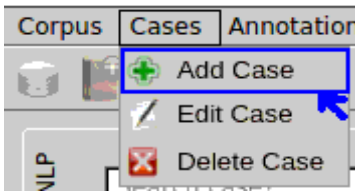
Resultado:



5b. **Click en el** *caso virtual 6*

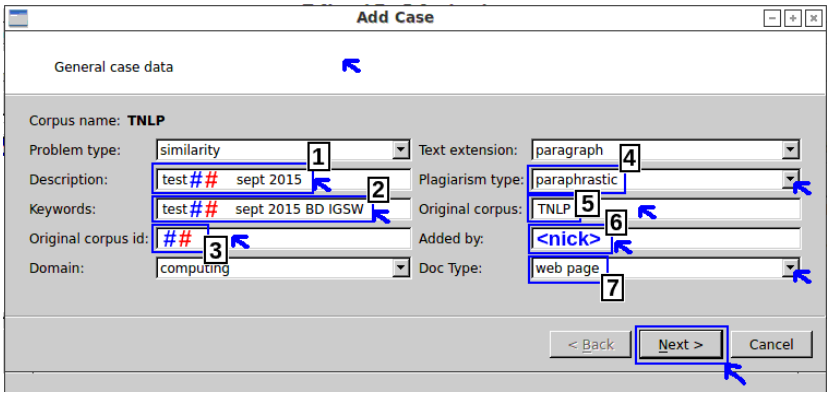
Add Case

Menú: Cases



a) Despliegue el menú "*Cases*", y seleccione la entrada "*Add Case".*

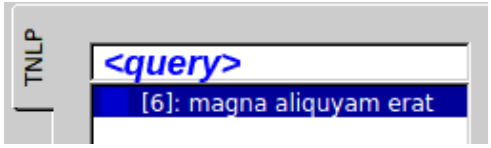
Wizard 1: Datos Generales del caso



b) Al apretar la entrada "Add Case" sale la 1ra parte del Wizard *"Add Case".*

Escribir y seleccionar cada opción:

1. El paso **Description**: test **# número asignado al colaborador**, **# número del caso que está elaborando en esos momentos**. Ejemplo: el colaborador **8** redacta su primer caso. Se sugiere como descripción: ***test81 sept 2015***.
2. **Keywords**: son las palabras que se usarán en el search de la aplicación para encontrar un caso específico en una lista larga. Ej.: ***test81 sept 2015 BD IGSW***

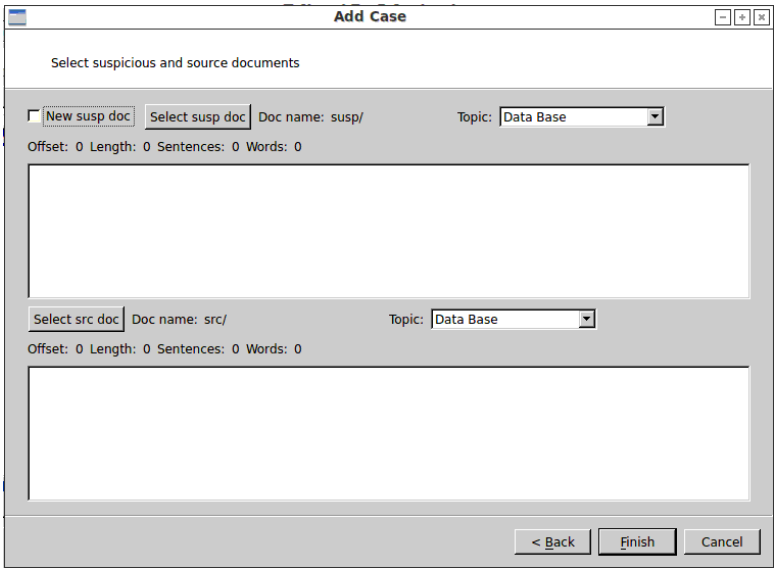


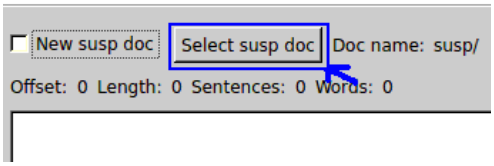
3. **Original Corpus ID**: es un id útil para preservar el orden y origen del caso. TNLP puede contener casos de corpus como el PAN o el P4P. Ej.:***81***.
4. **Plagiarism type**: permite identificar el tipo de texto-reusado fundamental en el caso. La mayoría de los casos de esta versión son **paraphrastic**, que significa *parapráticos o parafraseados del original*. El resto de los casos son **literal** o **none**.
5. **Original Corpus**: permite, similar a Original Corpus ID, permite preservar el origen del caso haciendo referencia al nombre del corpus del que proviene. Todos los casos de esta versión llevarán el nombre de **TNLP**.
6. **Added by**: nick o nombre personal del que introduce el caso, se recomienda que un mismo autor utilice siempre la misma seña. Ej.: **abelm**.
7. **Document Type**: en el caso de los textos que se utilizarán para esta versión son términos sobre computación extraídos de la wikipedia en español, se debe utilizar la opción **Web page**.

Wizard 2

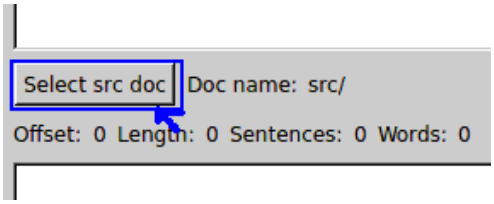
Seleccionar documentos del caso

c) al apretar el botón de “Next”, una vez llenados los parámetros anteriores aparece la segunda ventana del wizard “Add Case”.





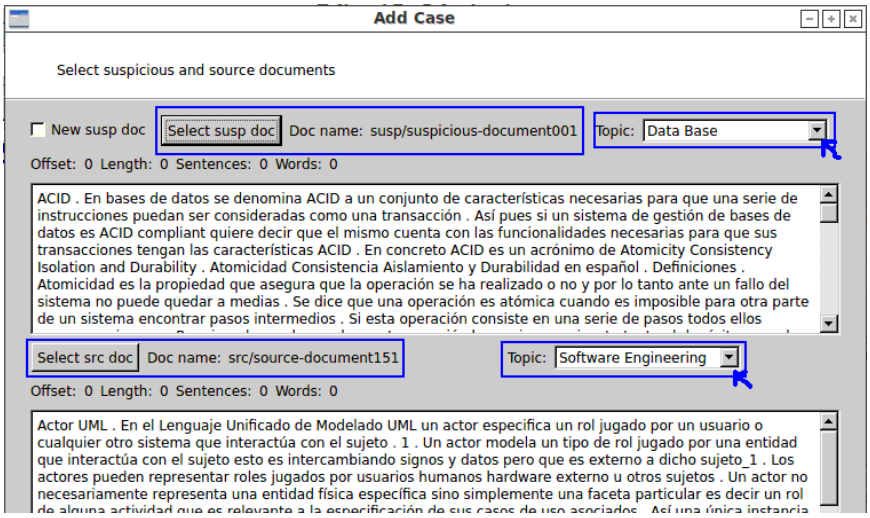
c.1) El siguiente paso es escoger el documento sospechoso. Ej.:**suspicious-document001.txt**.



c.2) Luego se escoge el documento fuente u original, aquel donde se encuentra el fragmento que ha sido plagiado en el documento sospechoso. Ej.:**source-document151.txt**.

Nota: ambos textos pueden ser escogidos aleatoriamente, aunque tanto susp como src con el mismo número son iguales.

Los textos pueden ser escogidos aleatoriamente. En esta guía brindamos una noción de la composición de todos los textos. En el grupo 1 los doc001 – 075 son sobre Bases de Datos, del 151 – 225 son sobre Ingeniería de Software. En el grupo 2 los doc del 076 – 150 son del tema Inteligencia Artificial, y del 225 – 300 son sobre Diseño y Test de Algoritmos. En la sección Indicaciones para crear los casos se especifica si los casos deben ser intra-topics or inter-topics.



c.3) Una vez escogidos los textos automáticamente se actualizan los *“Doc name”* en el wizard.

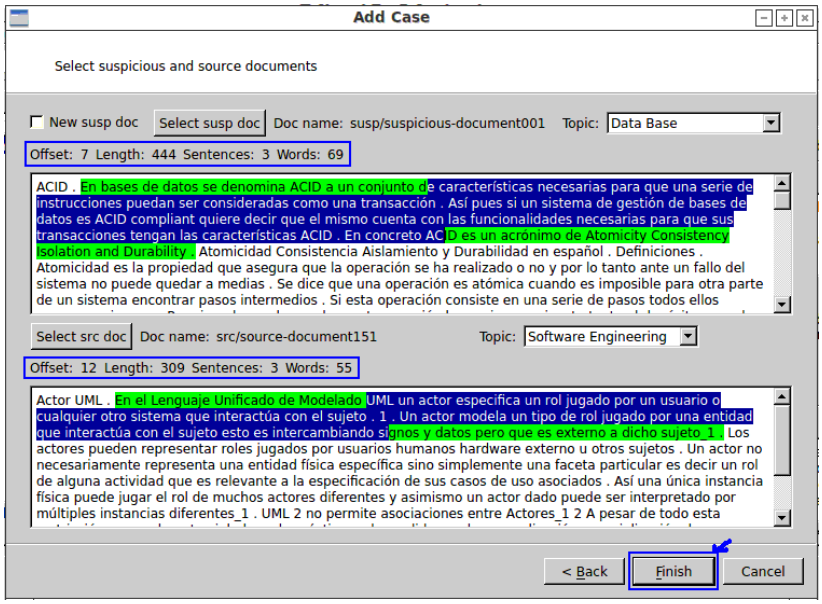
c.4) Luego se deben escoger los tópicos de cada documento. Ej.: **Data Base** y **Software Engineering**.

Seleccionar fragmentos del caso

d) Corresponde luego seleccionar los fragmentos del caso, o sea parte XXX del texto susp ≈ a la parte YYY del texto src.

Localice las oraciones consecutivas que estarán involucradas en ambos cuadros de texto utilizando el scroll.

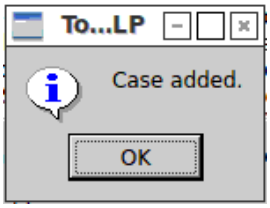
Haga click en cualquier punto de la primera oración involucrada y arrastre el mouse hasta cualquier punto de la última.



Resultado: las oraciones escogidas se tornarán verdes.

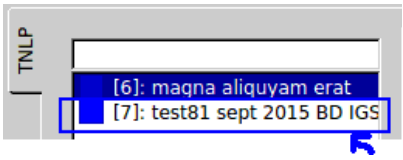
Note como los parámetros *Offset, Length, Sentences, Words* cambian a medida que usted selecciona. En la sección Indicaciones para crear los casos se recomiendan valores del número de *palabras* o *words*, para construir cada caso.

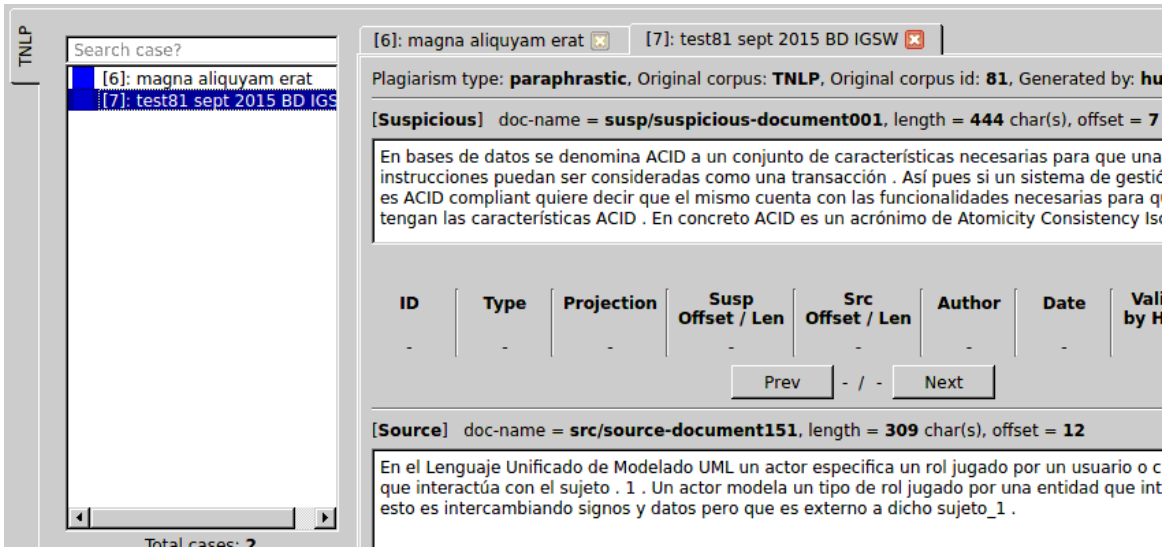
d.1) Al hacer click en el *"Botón Finish"* el caso será agregado al xml. Un mensaje lo indica en pantalla.



Resultado: se agrega un caso en la lista de QtNLP.

d.2) Se puede visualizar el caso haciendo click en el nuevo elemento de la lista.





Componer fragmento susp del caso

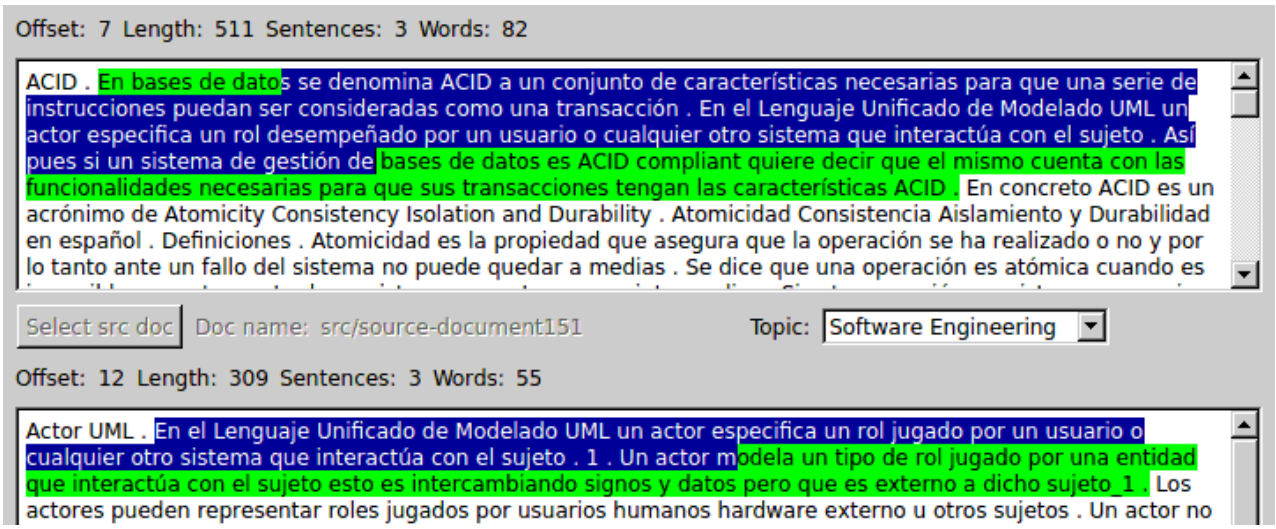
Una vez agregado un caso, repita todos los pasos anteriores, y deténgase en el paso d), justo antes de seleccionar los fragmentos.

Intente escribir en el cuadro de texto del elemento src o fuente. ¡Imposible!

Intente escribir en el cuadro de texto del elemento susp o sospechoso. ¡Vualá! Usted puede cambiar palabras en el fragmento identificado o incluso componer oraciones o párrafos completos. También puede copiar y pegar textos de otro lugar.

Note que puede seleccionar fragmentos en el src y copiarlos al susp. En algún momento esto será sugerido para elaborar cierto tipo de casos.

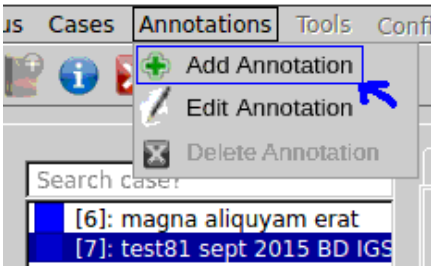
Hagamos un ejemplo de *editar una oración* para presentar la última funcionalidad de esta versión 1.0 de QtNLP, la adición de anotaciones.



En el fragmento susp se ha agregado, como 2da oración, la 1ra del fragmento src: “En el lenguaje Unificado de Modelado...”.
Sin embargo se ha cambiado la palabra *“...un rol **jugado**...”* por *“un rol **desempeñado**”*.
Adicionemos ahora una anotación a partir de este cambio. Ver sección Add Annotation.

Add Annotation

Menú:l Annotation

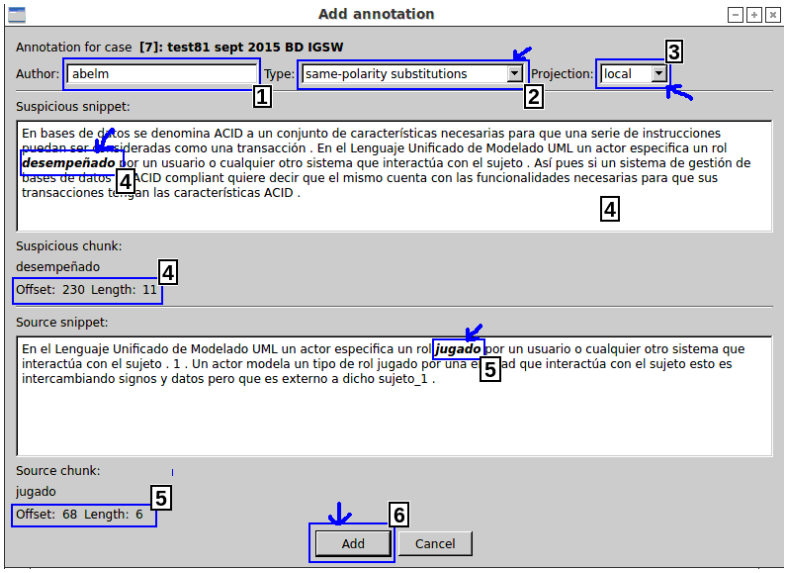


a) Despliegue el menú “Annotations”, y seleccione la entrada “*Add Annotations”*.
Las anotaciones que agreguen se harán al caso que esté en la pestaña que se encuentre en el foco. En este caso al caso 7 agregado.

Nota: en esta versión de QtNLP el Delete Annotation no está implementado.

Seleccionar palabras involucradas y datos

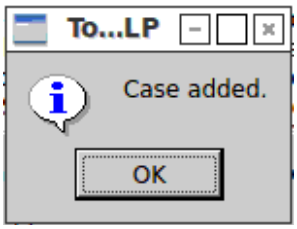
b) Al apretar la entrada “Add Annotation” sale el wizard *“Add Annotation”.*



Escribir y seleccionar cada opción:

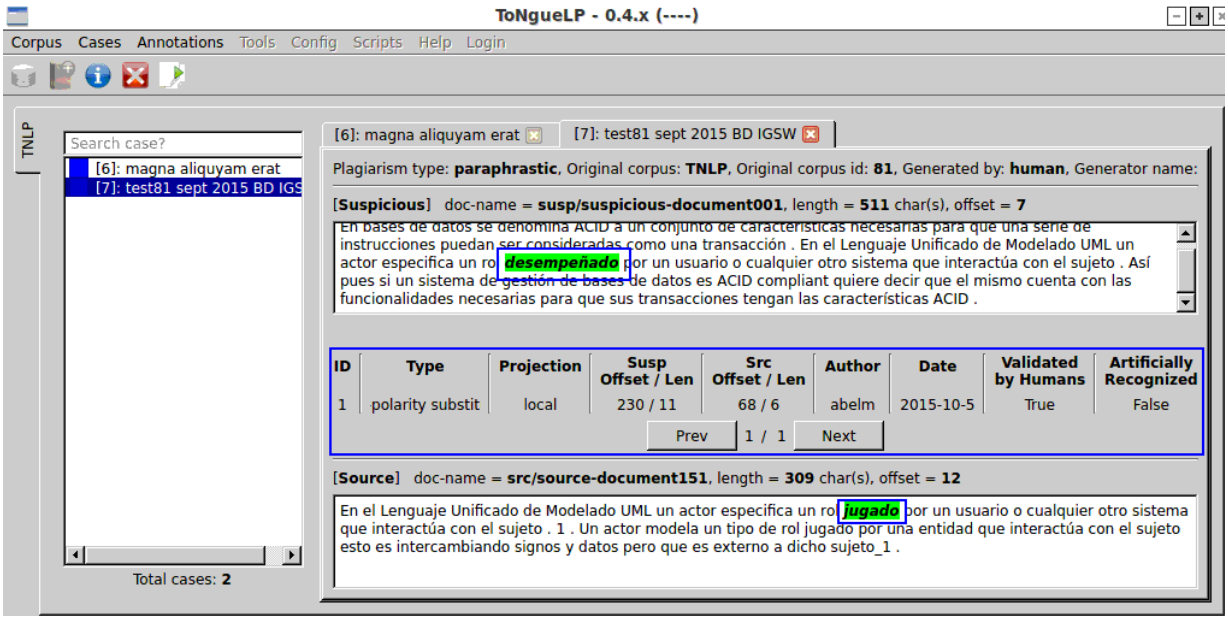
1. El paso **Author**: se debe agregar el mismo nick utilizado durante la creación del caso.
2. **Type**: refiere el tipo de paráfrasis utilizada, en el ejemplo actual es un sinónimo o **same polarity substitution**.
3. Projection: si el cambio que se está anotando obliga otras paráfrasis es **global** de lo contrario **local**.
4. Haga click sobre la palabra o la frase que conforman la anotación en el documento sospechoso. Notará como la palabra cambia a *negritas e itálicas*. Al mismo tiempo los datos *Offset y Length* del fragmento susp aparecerán automáticamente calculados.
5. Ibídem a 4 pero para el cuadro de texto src.
6. Al terminar de definir la anotación apriete el botón Add.

c) Al hacer click en el *“Botón Add”* la anotación será agregada al xml. Un mensaje lo indica en pantalla.



El Wizard de *Add Annotation* no se cerrará a la espera de más anotaciones para el mismo caso. Para cerrar el wizard utilice el botón “Cancel” o cierre el wizard.

El resultado puede verse como sigue:



La información de las anotaciones se ha actualizado. Y las palabras implicadas en la annotación que se está visualizando se resaltan en verde, similar a los fragmentos en el wizard de *“Add Case”*.

Si la lista de anotaciones crece puede usar los botones *“Prev”* y *“Next”* para visualizar otras anotaciones del mismo caso en el *área de información de las anotaciones*.

Referencias Bibliográficas

[1] M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso, “An Evaluation Framework for Plagiarism Detection,” in *Coling 2010, 23rd International Conference on Computational Linguistics*, 2010, no. August, pp. 997–1005.