

Arquitectura de información

22 de diciembre de 2014

Proyecto: ToNgueLP, editor de corpus de textos para tareas NLP

Producto: ToNgueLP Corpus Tools

Versión: 1.0

Autores: Abel Meneses Abad
Leonel Salazar Videaux

Institución: Cooperativa CubaSWL

Control de versiones

Fecha	Versión	Iteración	Descripción	Autor
29/6/2014	0.1	1	Digitalizada el prototipo de la vista principal con Pencil.	Abel Meneses Abad
Aug/2014	0.2	2	Cambios a partir de la discusión con el programador de UI.	Abel Meneses Abad
Sept/2014	0.3	3	Cambios en la UI Ppal tras análisis y diseño de procesos.	Abel Meneses Abad

Reglas de Confidencialidad

Documento parcialmente libre. Desarrollado para la tesis de doctorado “Análisis de relaciones semánticas para la detección de plagio parafrástico en textos científicos sobre computación escritos en idioma español”, del autor Abel Meneses Abad. El mismo constituye evidencia del desarrollo de ToNgueLP herramienta parcialmente libre desarrollada para lingüistas y con el apoyo de la Universidad de Camagüey, y la Universidad Marta Abreu de Las Villas, así como la asesoría del PhD Alberto Barrón Cedeño. Factible de exportar por las partes titulares.

Este documento contiene información propietaria de [Abel Meneses Abad](#), y es emitido confidencialmente para orientar al equipo de desarrollo de ToNgueLP. Su utilización en entornos académicos o de investigación sin fines comerciales es permitido, así como su reutilización literal en documentos afines, citando el original y sin fines de lucro. Si elabora un artefacto con fines comerciales debe acordar la compra del original con los titulares del producto.

El que recibe el documento asume la custodia y control, comprometiéndose a no reproducir, divulgar, difundir o de cualquier manera hacer de conocimientos público su contenido, excepto para cumplir el propósito para el cual se ha generado.

El documento ha sido elaborado sobre la plantilla de Arquitectura de Información de SXP 4.x, autoría y titularidad de Enma Lidia Muñoz García.

Estas reglas son aplicables a las **11** páginas de este documento.

Índice

1. Introducción	4
1.1. Alcance	4
1.2. Definiciones, Acrónimos y Abreviaturas	4
1.3. Referencias	4
2. Esbozo de la estructura o taxonomía	4
2.1. Descripción de los elementos de la arquitectura	5
2.2. Definición de los elementos de usabilidad y accesibilidad	6
3. Definición de la estructura	8
3.1. Diagrama de arquitectura	8
3.2. Elementos del sistema de navegación	8
3.3. Diseño de las pantallas tipo	10
3.3.1. Vista Principal	10
3.3.2. Vista de Comparación (Borrador)	13
3.3.3. Vista de Diccionarios (Borrador)	14
3.4. Descripción de la interacción de los usuarios	15
4. Componentes relevantes del diseño	15

1. Introducción

Este documento define la arquitectura de información de ToNgueLP.

1.1. Alcance

Este documento refiere en su versión 1 los elementos de la arquitectura de información para las 3 vistas fundamentales de ToNgueLP:

- Vista Principal.
- Vista de Comparación.
- Vista de Diccionarios.

1.2. Definiciones, Acrónimos y Abreviaturas

1.3. Referencias

1. Estudio de Homólogos del proyecto ToNgueLP.
2. Ayuda generada con Sphinx del proyecto ToNgueLP.
3. Notas manuscritas de la tesis de doctorado de Abel Meneses Abad.
4. Lista de chequeo del estándar ISO 9241-210:2010
5. Diagrama_de_Estados_Vista_Principal_ToNgueLP-Corpus_Process.dia
6. Diagrama_de_Estados_Vista_Principal_ToNgueLP-Case_Process.dia

2. Esbozo de la estructura o taxonomía

1. Corpus
 - a) Load New Corpus
 - b) Insert New Doc
 - c) Save Corpus
 - d) Corpus Inf.
 - e) Select Parser
2. Tools
 - a) Vista Principal

- b) Vista Comparación
- c) Vista de Diccionario

3. Config

- a) Config Lang
- b) Config Parser
- c) Config Paths
- d) Config User Data

4. Scripts

- a) (nombre x definir): Convertir documentos TXT o PDF a ZZZdoc.XML.
- b) Verificar compatibilidad del Corpus cargado con el Parser configurado.
- c) (nombre x definir): Generar XML estandar a partir de la carpeta OUT de la salida de un algoritmo de detección.
- d) (nombre x definir): Comparación del XML estandar con XML del corpus cargado.
- e) (nombre x definir): Actualizar ciertos tags del XML de datos del algoritmo y actualizar ciertos tags del XML reporte de casos detectados.

5. Help

- a) Copyright
- b) Ayuda de Usuarios
- c) Manual de Desarrollo
- d) About
- e) Contact us

6. Users

- a) Loguin

2.1. Descripción de los elementos de la arquitectura

1. “*Barra de Menús*”: Barras de Menú
2. “*Loguin*”: Opción de Loguin, para expertos autorizados a confirmar manualmente los casos generados automáticamente.
3. “*Tabs Corpus*”: Acceder simultáneamente a diferentes corpus cargados.

4. “*Barras de Herramientas*”: Barra de Herramientas de Menús.
5. “*Search*”: Componente de búsqueda.
6. “*Case List*”: Lista de casos
7. “*Tabs Case*”: Tabs de casos y/o vistas de la aplicación.
8. “*src Case Text*”: Fragmento del texto del documento fuente u original.
9. “*src Case Bar Data*”: Barra de datos del fragmento fuente.
10. “*susp Case Text*”: Fragmento del texto del documento sospechoso.
11. “*susp Case Bar Data*”: Barra de datos del fragmento sospechoso.
12. “*Barra de Edición de Casos*”: Elementos para editar fragmentos.

<<Descripción textual de cada uno de los elementos de la estructura, características, comportamiento. etc.>>

2.2. Definición de los elementos de usabilidad y accesibilidad

1. “*Barra de Menús*”: No se desplegará, sus contenidos serán visualizados en [4, Barra de Herramientas]. El objetivo es conservar la visión de barras desplegables, pero eliminarlas, visualmente el efecto se verá, y se gana en el tamaño de los elementos a clicar.
2. “*Loguin*”: No se levanta una ventana sino un menú embebido como hacen todas las aplicaciones web actuales.
3. “*Tabs Corpus*”: Similar al comportamiento de tabs laterales de konqueror,... Acceder simultáneamente a diferentes corpus cargados.
4. “*Barras de Herramientas*”: click o movimiento a la derecha o izquierda si hubieran muchas opciones o elementos a clicar.
5. “*Search*”: Al comenzar a tipear y pasar de 3 caracteres automáticamente se comienzan a mostrar resultados, desplegándose por debajo del componente y por encima de la interfaz sin formar parte de ella. Similar a componentes de búsqueda existentes en otras aplicaciones como Kiwix.
6. “*Case List*”: En este elemento solo se puede dar click sobre sus items.
7. “*Tabs Case*”: Similar a los tabs de Firefox y navegadores de archivo. Se pueden visualizar módulos completos de la herramienta y no solo los casos.
8. “*src Case Text*”: Elemento de texto que posee posibilidades de recargarse como si fuera AJAX, teniendo guardados, pero invisibles, elementos como las “*annotations*”.
9. “*src Case Bar Data*”: elemento simple que se autorecargan los valores al hacer cualquier modificación del elemento [8]

10. “*susp Case Text*”: Similar a [8]
11. “*susp Case Bar Data*”: Similar a [9]
12. “*Barra de Edición de Casos*”: Similar a [4] algunos elementos cambiarán sus labels al clickearlos en dependencia del estado en el que esté la aplicación, permitiendo una nueva función asociada con la anterior. Esto permite ahorrar espacio, facilitar al usuario la culminación satisfactoria de una instancia de cualquier proceso relacionado con un caso. Al mismo tiempo el usuario verá las opciones en un elemento de gran tamaño.

3. Definición de la estructura

3.1. Diagrama de arquitectura

Diagramación del producto

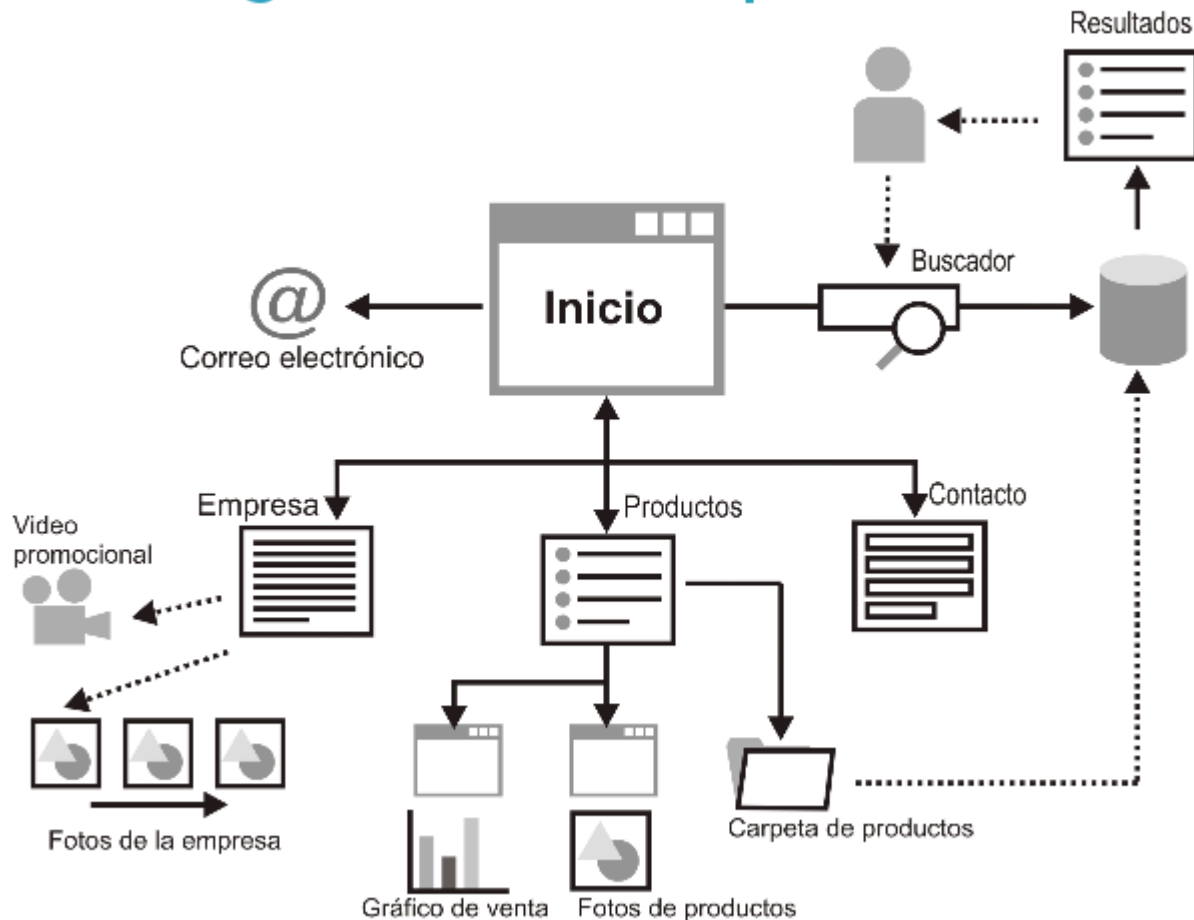


Figura 1: Ejemplo de Diagrama. Rodrigo, 2009. La Habana.

3.2. Elementos del sistema de navegación

Aquí se definen los elementos globales que podrán accesarse en todas las vistas, y además la conceptualización de algunos de ellos que es fundamental para el entendimiento de la usabilidad de la aplicación.

1. Elemento [1] “*Barra de Menús*”: Barras de Menús (definición beta)→ Estas no se desplegarán serán visualizadas

en el elemento “*Barras de Herramientas*”.

- a) Menú “*Corpus*”: Permite cargar un archivo de corpus/ salvar el corpus en edición/ o todos los corpus.
 - 1) Load New Corpus
 - 2) Insert New Doc
 - 3) Save Corpus
 - 4) Corpus Inf.
 - 5) Seleccionar el Parser
 - 6) Sugerir automáticamente el parser compatible
 - b) Menú “*Tools*”: Vista Principal/Vista Comparación/Vista de Diccionario. (el futuro puede tener las otras vistas previstas en la LRP)
 - c) Menú “*Config*”: Permite definir el Idioma/Parser/Datos del Editor/... (falta bien esta por estudiar, depende mucho de las cosas a programar.)
 - d) Menú “*Scripts*”: Acceso directo a los procesos de ToNgueLP para ejecutarlos manualmente o ver la salida que van generando(similar a la consola de synaptic cuando está instalando):
 - 1) (nombre x definir): Convertir documentos TXT o PDF a ZZZdoc.XML.
 - 2) Verificar compatibilidad del Corpus cargado con el Parser configurado.
 - 3) (nombre x definir): Generar XML estandar a partir de la carpeta OUT de la salida de un algoritmo de detección.
 - 4) (nombre x definir): Comparación del XML estandar con XML del corpus cargado.
 - 5) (nombre x definir): Actualizar ciertos tags del XML de datos del algoritmo y actualizar ciertos tags del XML reporte de casos detectados.
 - 6) ~~(el futuro) “Adicionar nuevo script”, tener los elementos posibles en pantalla y unirlos como en un simulador o sea moviendo cuadritos(preguntar a Leonel como se llamaban los elementos visuales desplazables que había en el UI del SCADA).~~
 - e) Menú “*Help*”: enlaces a Copyright/Ayuda de Usuarios/Manual de Desarrollo/About.
2. Elemento [2] “*Loguin*”: Opción de Loguin para expertos autorizados a confirmar manualmente los casos generados automáticamente.
 3. Elemento [3] “*Tabs Case*”: Tener acceso a más de un corpus simultáneamente.
 4. Elemento [4] “*Barras de Herramientas*”: Su funcionamiento es similar al componente “Barras de Herramientas” de Microsoft Office 2010.¹
 5. Elemento [5] “*Search*”: Componente de búsqueda: buscar casos que contienen determinada secuencia de palabras (aquí se debe agregar un selector con: id, corpus, case, query)
 6. Map Doc Bar [...] : Permite acceder a diferentes posiciones del documento marcados con algún objetivo. Puede ser marcas que indiquen fragmentos re-usados de texto.

3.3. Diseño de las pantallas tipo

Se describen 3 pantallas tipos:

- Vista Principal.
- Vista de Comparación.
- Vista de Diccionarios.

3.3.1. Vista Principal

Esta pantalla es la pantalla inicial de ToNgueLP, su objetivo principal es visualizar los casos contenidos en los XML de los corpus contenidos por la aplicación.

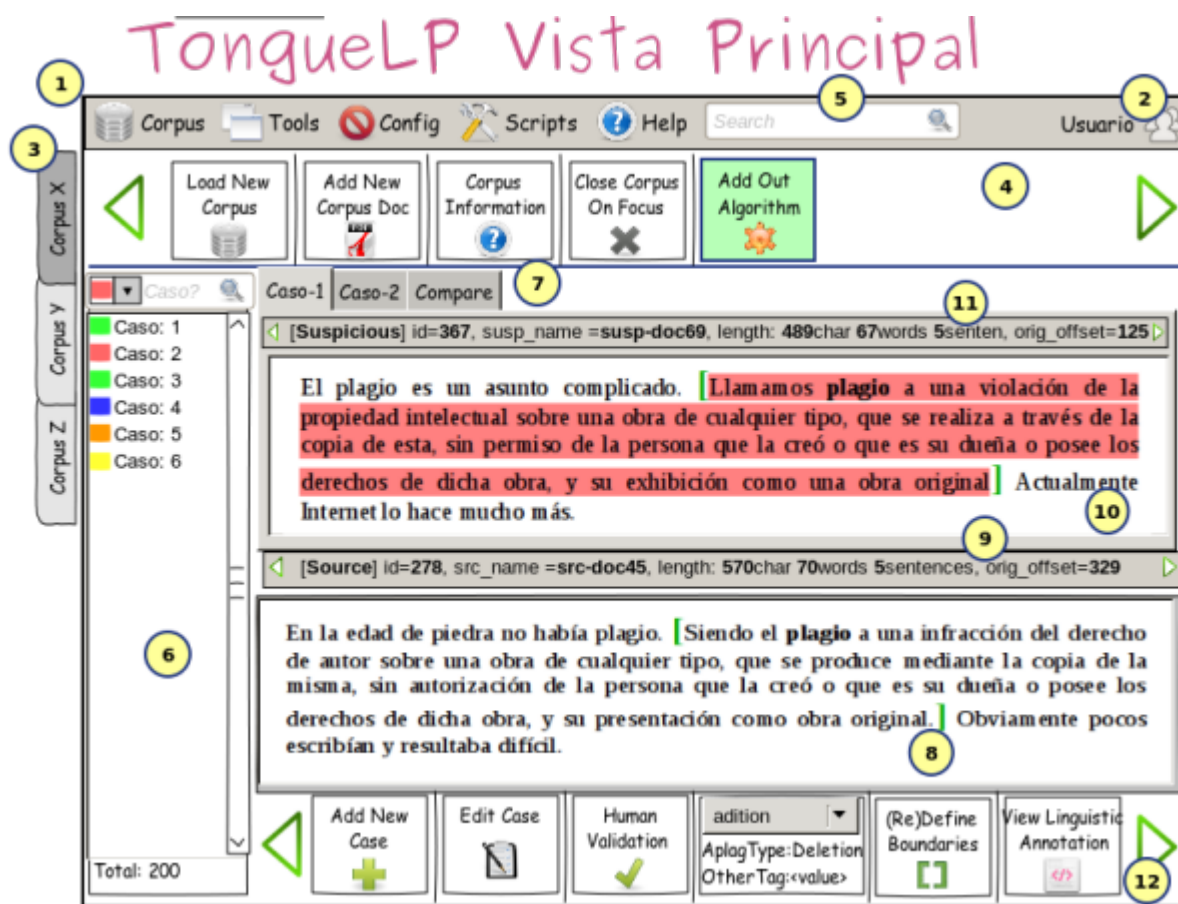


Figura 2: Vista Principal ToNgueLP

- Breve descripción de los elementos que componen la pantalla:

1. “*Barra de Menús*”: Barras de Menú
2. “*Loguin*”: Opción de Loguin, para expertos autorizados a confirmar manualmente los casos generados automáticamente.
3. “*Tabs Corpus*”: Acceder simultáneamente a diferentes corpus cargados.
4. “*Barras de Herramientas*”: Barra de Plugins o .
5. “*Search*”: Componente de búsqueda.
6. “*Case List*”: Lista de casos
7. “*Tabs Case*”: Tabs de casos y/o vistas de la aplicación.
8. “*src Case Text*”: Fragmento del texto del documento fuente u original.
9. “*src Case Bar Data*”: Barra de datos del fragmento fuente.
10. “*susp Case Text*”: Fragmento del texto del documento sospechoso.
11. “*susp Case Bar Data*”: Barra de datos del fragmento sospechoso.
12. “*Barra de Edición de Casos*”: Elementos para editar fragmentos.

■ **Descripción detallada de los elementos que componen la pantalla:**

- Elemento [6] “*Case List*”: Mostrará en una lista todos los casos existentes en el corpus para accederlos rápidamente(en caso de que sean muchos este componente deberá tener también una búsqueda para filtrarlos y accederlos rápidamente).
 - El “Search Case List” será un elemento de alto valor agregado pues: permitirá buscar por los colores que identifican los tipos de texto re-usado.
 - ◇ Permitirá buscar por palabras claves asociadas al texto que contienen los fragmentos asociados.
- Elemento [7] “*Tabs Case*”: Tabs de, casos o vistas de la aplicación: casos en edición, o Vistas abiertas en las que se trabaja simultáneamente.
- Elemento [8] “*src Case Text*”: Fragmento del texto del documento fuente u original: visualiza el fragmento original del caso (este componente no permite edición al menos que se active el ELEMENTO “Active Text Edition” después de autenticado) - Este componente debe permitir el formateado de texto: textos en colores, textos subrayados, click derecho para adicionar al diccionario, etc.
- Elemento [9] “*src Case Bar Data*”: Barra de datos del fragmento fuente: datos que contiene el corpus sobre el fragmento original.(no se edita manualmente, solo automáticamente).
- Elemento [10] “*susp Case Text*”: Fragmento del texto del documento sospechoso: visualiza el fragmento sospechoso del caso (este componente no permite edición al menos que se active el ELEMENTO “Active Text Edition” después de autenticado) - Este componente debe permitir el formateado de texto: textos en colores, textos subrayados, click derecho para adicionar al diccionario, etc.
- Elemento [11] “*susp Case Bar Data*”: Barra de datos del fragmento sospechoso: datos que contiene el corpus sobre el fragmento sospechoso.(no se edita manualmente, solo automáticamente)

- Elemento [12] “*Barra de Edición de Casos*”: Funciones para editar fragmentos:
 - Elemento “*Add New Case*”: Adicionar un nuevo caso. Se colocan en blanco 8 y 10. (un algoritmo(basado en análisis temático) sugiere 3 documentos donde serán insertados, se abre un popup para seleccionar y estos son insertados automáticamente).
 - Elemento “*Edit Case*”: Permite editar 8 y 10 en el caso de la pestaña en el foco.
 - ◊ “*Verify Case*”: Segundo estado de “*Edit Case*”, una vez editado necesita verificarse.²
 - ◊ “*Update Case*”: Tercer estado de “*Edit Case*”, una vez verificado puede insertarse los cambios en el XML del corpus.³
 - Elemento “*Human Validation*”: Para que se confirme manualmente que el caso generado por un algoritmo automático realmente es consistente.
 - ◊ “*Save Validation*”: Segundo estado de “*Human Validation*”, una vez llenados los campos en blanco de “*Plag Type*” puede ser guardado como datos del caso.⁴
 - Elemento “*Plag Type*”: Datos generados automáticamente sobre el caso. Nota: se activa para edición y un selector para que manualmente se categorice a este caso.
 - ◊ Contiene los siguientes datos: Clase de Plagio, Tipo de Plagio, Projection.
 - Elemento “*Re-Define Boundaries*”: Permite definir o redefinir los límites (denotado como corchetes verdes) de los textos que componen el caso. Siempre podrá verse texto que no pertenece al caso, en los elementos “*src Case Text*” y “*susp Case Text*” por fuera de los límites superior e inferior.
 - Elemento “*View Linguistic Annotation*”: Permitirá adicionar etiquetas o visualizar aquellas que existen en el corpus, como: POS, SRL, Name Entity, etc.
 - ◊ “*Next Annotation*”: Segundo estado de “*Corpus Annotation*”, permitirá ver el siguiente tipo de anotación contenido, ej.: POS, SRL, etc.⁵

3.3.2. Vista de Comparación (Borrador)

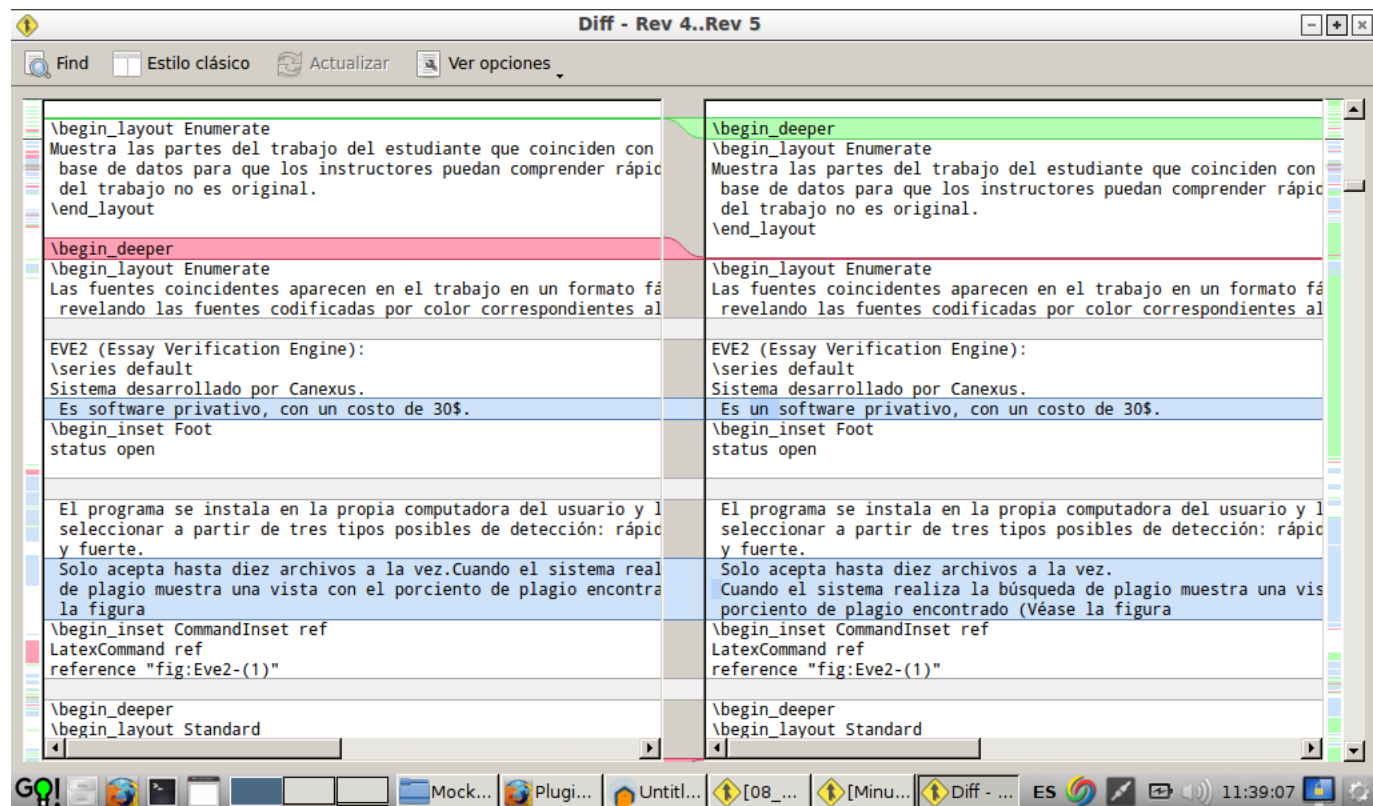


Figura 3: Vista de Diff de Bazaar Explorer, prototipo de como quedará la Vista de Comparación.

- Breve descripción de los elementos que componen la pantalla:
- Descripción detallada de los elementos que componen la pantalla:

3.3.3. Vista de Diccionarios (Borrador)

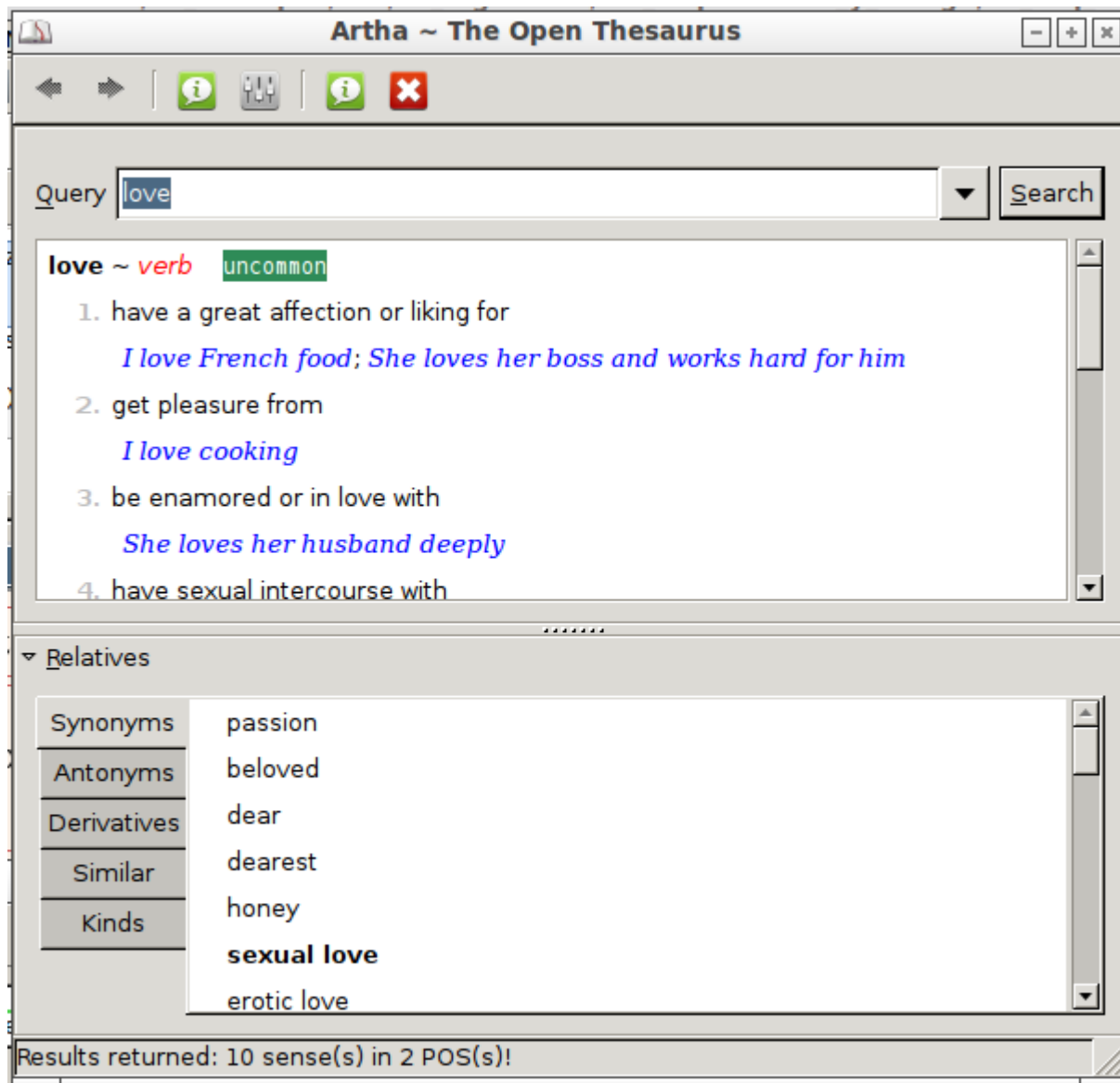


Figura 4: Vista del Front-End de Wordnet: Artha.

- Breve descripción de los elementos que componen la pantalla:
- Descripción detallada de los elementos que componen la pantalla:

3.4. Descripción de la interacción de los usuarios

<<Debe explicarse la interacción de los elementos y los resultados en sus pantallas correspondientes.>>

4. Componentes relevantes del diseño

<<Deben listarse y explicar los componentes relevantes que se incluirán en el diseño con su nombre técnico y su correspondiente referencia internacional.>>

<<La idea de este punto es ir aumentando el acervo del equipo de desarrollo en cuanto a diseño de funcionalidades de acuerdo a lo más novedoso usado actualmente.>>

Notes

¹Similar a Doclux, permite acceder directamente las funciones principales de la aplicación.

²El documento “Diagrama de Estados Vista Principal de ToNgueLP refleja las condiciones de este estado.”

³El documento “Diagrama de Estados Vista Principal de ToNgueLP refleja las condiciones de este estado.”

⁴El documento “Diagrama de Estados Vista Principal de ToNgueLP refleja las condiciones de este estado.”

⁵El documento “Diagrama de Estados Vista Principal de ToNgueLP refleja las condiciones de este estado.”