

ULISSES: un *Integrated Development Environment* desarrollado para la anotación de un *corpus* romancístico

Natália Albino Pires
Escola Superior de Educação de Coimbra
Praça dos Heróis do Ultramar, s/n
3030-329 Coimbra
npires@esec.pt

Resumen: Ulisses es un *Integrated Development Environment* desarrollado para la anotación de un *corpus* constituido por romances de la tradición oral moderna portuguesa y que tiene la particularidad de poseer una estructura modular que admite la integración de nuevas herramientas y funcionalidades.

Palabras clave: Léxico; *Corpus*; Romancero; Ulisses; IDE

Abstract: Ulisses is an Integrated Development Environment developed for the annotation of a *corpus* comprised of ballads from the portuguese modern oral tradition, featuring a modular structure which allows the integration of new tools and functionality.

Keywords: Lexicon, *Corpus*, Ballads; Ulisses; IDE

1 Introducción

La aplicación que ahora presentamos se desarrolló para el estudio del léxico de un *corpus* constituido por 1721 textos, versiones del romancero de la tradición oral moderna portuguesa editadas entre 1828 y 1960. La especificidad de nuestra aplicación resulta, sobre todo, de la especificidad de nuestro *corpus*¹ y de los objetivos que nos hemos propuesto para nuestra tesis de doctorado: definir el léxico del *corpus* y verificar si, a partir del léxico principal, se pueden encontrar léxicos particulares de informante, léxicos particulares de romances, de temas, de regiones o de editor, ya que algunos de los editores, sobre todo en el siglo XIX, han intentado darles a los textos un carácter más culto “corrigiendo” rasgos tradicionales.

2 Descripción y arquitectura de la herramienta

Ulisses se presentó por primera vez en el VII Congrès de Linguística General que se realizó en Barcelona entre el 18 y el 21 de

abril del 2006 y se caracteriza, sobre todo, por ser un IDE (*Integrated Development Environment*), que proporciona una *interface* versátil y sofisticada y que reúne bajo un único Ambiente de Trabajo todas las herramientas que le permiten al investigador introducir, editar, catalogar, anotar, procesar y analizar *corpora*. Anclado en una estructura modular, admite la integración de nuevas herramientas o funcionalidades que no hayan sido contempladas originalmente, pudiendo, por lo tanto, adaptarse fácilmente para corresponder a las exigencias específicas de una determinada área de investigación.

En lo que se refiere a pormenores técnicos, se desarrolló en C#, tiene como motor de base de datos el SQLite, requiere el .NET Framework 2.0, necesita el Windows XP e importa y exporta la información y la metainformación del *corpus* en formato XML.

Posee un *tokenizer* y un *tagger* automáticos, estando ambos basados en un algoritmo muy simple ya que el objetivo de nuestra investigación no se centra en desarrollar un *tagger* o un *tokenizer*, sino en estudiar el léxico del *corpus* con el auxilio de una aplicación informática.

¹ Nuestro *corpus* contiene versiones de cerca de cien romances distintos de la tradición oral moderna portuguesa, recopilados en diferentes regiones y editados por distintos editores a lo largo de 132 años.

No obstante, el investigador puede optar por tokenizar los textos y etiquetar los tokens tanto automáticamente como manualmente. Además, se le permite al investigador repasar todos los textos de modo que pueda corregir los errores, sea de tokenización, sea de etiquetaje. Ulisses cuenta, incluso, con un buscador que le permite al investigador buscar en el *corpus* palabras *tokens*, lemas, *tags* y otro tipo de anotaciones.

Sin embargo, la aplicación presentada no está aún terminada y por consiguiente no se encuentra todavía disponible online.

Muy pronto contará con un módulo de análisis estadístico del *corpus*, incorporando el cálculo de medias, desviaciones, *chi-square*, frecuencias absolutas, frecuencias relativas y ocurrencias, y un módulo de presentación de los datos estadísticos en forma de cuadros y/o listas, con la posibilidad de imprimirlos o exportarlos para SPSS, Access o Excel.

3 Otras herramientas similares

Ulisses, en cuanto IDE y en su filosofía, se puede comparar al proyecto GATE (General Architecture for Text Engineering), que se puede consultar en <<http://gate.ac.uk/>>, y al proyecto Ellogon, que se puede consultar en <www.ellogon.org/>. No obstante, GATE no ofrece un *tagger* que reconozca el portugués y cuando empezamos el análisis de nuestro *corpus* Ellogon se estaba todavía desarrollando².

4 Descripción de la Demo

En la demostración, y a través de un vídeo de captura de imagen, enseñaremos las funcionalidades principales del programa. De este modo, se pueden seguir los pasos necesarios para la construcción y anotación de un nuevo *corpus*, ejemplificándose también todo el proceso, desde la creación del etiquetario, de los campos de catalogación y de anotación, a la apertura de un *corpus* de trabajo (parte de nuestro *corpus*) y su tokenización, su etiquetaje, su catalogación y la búsqueda de palabras o de *tokens* etiquetados, sea en el texto, sea en todo el *corpus*.

Empezaremos con la construcción del etiquetario en la ventana de *tag manager*, en

donde el utilizador puede definir la terminología lingüística que pretende adoptar, las etiquetas y sus relaciones jerárquicas, y los colores a atribuir (o no) a cada una de ellas. En esta ventana, el utilizador tiene, además, la posibilidad de añadir y eliminar etiquetas o de reorganizar el etiquetario cuando lo desee.

A continuación, pasaremos a la ventana de *attributes manager*, en donde el investigador puede determinar todos los campos de catalogación de los textos de su *corpus*, los cuales pueden contener informaciones textuales, numéricas, de *drop down list* y *boolean*. Como en la anterior ventana, al utilizador se le permite añadir o borrar los campos cuando lo desee.

Para terminar, entraremos en la ventana de *annotation manager*, en donde el utilizador puede crear, alterar o eliminar todos los campos de anotación complementarios para el análisis de su *corpus*. Ejemplificaremos con las anotaciones que hemos considerado fundamentales para nuestro estudio.

Ejemplificada la creación del etiquetario, de los campos de catalogación y de las anotaciones extratextuales necesarias, abriremos las ventanas de trabajo que nos permitirán tokenizar y etiquetar cada texto del *corpus*. Estas ventanas de trabajo se pueden colocar en la disposición que el utilizador considera más conveniente: minimizándolas, redimensionándolas u ocultándolas.

Terminaremos la demostración con ejemplos del proceso de etiquetaje y de anotación interactivo del texto y con ejemplos de las posibilidades de búsqueda en el texto y en el *corpus*.

4.1 Tiempo de la demo

Para la demostración de las funcionalidades de la aplicación necesitaremos aproximadamente 20 minutos.

² Los lenguajes de programación utilizados por GATE y Ellogon fueran también un *handicap*, ya que el informático que realizó el Ulisses no domina ni Java ni TeL, respectivamente.