Ulisses – un IDE (*Integrated Development Environment*) para la anotación y procesamiento de corpora

Natália Albino Pires

ESEC – Praça dos Heróis do Ultramar, s/n, 3030-329 Coimbra – Portugal

Abstract

In this paper, we intend to account for how the singularities of a corpus comprised of versions from the romancero of the portuguese modern oral tradition, totalling 1721 texts, have compelled us to build a new software application that answers the needs and goals of our research; afterwards we will describe the Ulisses application, the IDE which was specifically developed for the annotation and analysis of the corpus.

Resumen

En esta comunicación pretendemos dar cuenta de cómo las singularidades de un corpus constituido por versiones del romancero de la tradición oral moderna portuguesa, un total de 1721 textos, nos han obligado a desarrollar una nueva aplicación informática que respondiera a las necesidades y a los objetivos de nuestro trabajo de investigación; y, a continuación, describiremos la aplicación Ulisses, el IDE desarrollado específicamente para la anotación y análisis del corpus.

Palabras clave: romancero, léxico, anotación, corpus, Ulisses.

1. Introducción

En el ámbito de nuestra tesis doctoral nos hemos propuesto estudiar un *corpus* romancístico constituido por 1721 textos, que corresponden a versiones de más de cien romances de la tradición oral moderna portuguesa recopilados en distintas regiones y publicados por diferentes editores entre 1828 y 1960.

En nuestro trabajo de investigación hemos tenido como objetivo principal el estudio del léxico y la construcción de un diccionario/vocabulario de formas flexionadas con la indicación del romance/versión en el que aparece cada forma. De este modo, hemos necesitado anotar todo el *corpus* ya que la comprobación de la especificidad del léxico conllevaría el estudio de las categorías de palabras Sustantivo, Adjetivo, Verbo, Adverbio y Cuantificador.

En esta comunicación pretendemos, por un lado, dar cuenta de cómo la especificidad de nuestros objetivos y las singularidades de nuestro *corpus* nos han obligado a desarrollar una nueva aplicación informática que respondiera a las necesidades de nuestro trabajo de investigación y, por otro lado, pretendemos describir la aplicación Ulisses, el IDE (Ambiente Integrado de Desarrollo) desarrollado específicamente para la anotación y análisis del *corpus*.

1.1. La especificidad del corpus romancístico

Las singularidades de nuestro corpus se hallan en las características específicas del género literario romance: texto poético con tiradas de versos mayoritariamente monorrimos con rima

asonante y con una estructura sintáctica que presenta inversiones con las que, tanto los referentes como los actantes, surgen topicalizados.

A su vez, y porque "as regiões romancísticas não correspondem a entidades administrativas" (Araújo, 1998: 222), parte de su especificidad reside en el hecho de que los textos originarios de zonas fronterizas nos aparecen recitados en castellano, en gallego, en mirandés, en portugués, en castellano/portugués, en gallego/portugués, en castellano/mirandés, en mirandés/portugués o incluso en castellano/gallego/portugués y en castellano/mirandés/portugués. Por fin, otra de las características fundamentales del corpus estudiado parte del hecho de que está constituido por versiones de romances recopiladas en distintas épocas y regiones, editadas a lo largo de 132 años con diferentes criterios.

1.2. La anotación del corpus

Los rasgos que habíamos pretendido anotar en nuestro *corpus* se relacionan directamente con los objetivos definidos para nuestra investigación (la construcción de un diccionario/vocabulario y la comprobación de la especificidad del léxico) y con la especificidad del género literario objeto de estudio.

De este modo, para que pudiéramos cumplir nuestros objetivos, se volvió imprescindible:

i. mantener informaciones extratextuales como el origen de los textos, el nombre de su primer editor y la fecha de su primera edición, un código de referencia para saber si el texto está o no contaminado, un código de versión, un código de tema/asunto y un código de romance.

ii. atribuir una anotación a cada *token* en fin de verso.

iii. indicar la lengua a la que pertenece determinado *token*, dando cuenta, si es el caso, de su ambigüedad porque hay palabras que gráficamente son iguales en todas las lenguas en contacto en los textos: portugués, gallego, castellano y mirandés.

iv. una herramienta de *full tagging*.

2. Búsqueda de herramientas y aplicabilidad de los analizadores morfológicos al *corpus* romancístico

En los últimos diez años, ha aumentado significativamente el número de proyectos y de aplicaciones en el ámbito del Procesamiento de Lenguage Natural (PNL); no obstante y que sepamos, hasta el presente se han desarrollado solamente cuatro analizadores morfológicos o morfosintácticos para el Portugués Europeo (PE): Jspell[1], Palavroso[2], VISL[3] y LX-Suite[4].

Antes de optar por el diseño de una nueva aplicación, aceptando todos los riesgos que tal decisión conlleva, hemos contactado con los directores de tres de los proyectos y hemos testado con un texto de muestra cada uno de los analizadores disponibles para el PE. Por

[1] Proyecto elaborado en la Universidade do Minho e integrado en el Projecto Natura.

[2] Desarrollado por el Grupo de Linguagem Natural de INESC y funciona actualmente como corrector ortográfico.

[3] Se trata de un proyecto desarrollado en Noruega.

[4] Desarrollado por un equipo conjunto de miebros del Departamento de Informática de la Faculdade de Ciências de la Universidade de Lisboa y por miembros del Centro de Linguística de la Faculdade de Letras de la Universidade de Lisboa.

problemas personales, el director del proyecto Jspell no pudo seguir ayudándonos y aunque los directores de los demás proyectos mostraran total disponibilidad para ayudarnos con el etiquetaje de los textos, problemas técnicos nos impidieron utilizarlos. Tanto LX-Suite como VISL se diseñaron para que fueran utilizados por los lingüístas y, por lo tanto, no permiten mantener informaciones extra-textuales que, en nuestro caso particular, son fundamentales para crear *subcorpora* y comparar los datos sobre el léxico. Además, LX-Suite nos presenta otro problema: se trata de una herramienta de "shallow processing of portuguese" y no atribuye información morfosintáctica a los tokens.

A partir de un test hemos comprobado otro problema que nos impidiría la aplicación de uno de los analizadores a nuestro *corpus*: cada uno de los tres analizadores disponibles para el PE posee un diccionario preparado para reconocer únicamente el portugués y, aunque estudiáramos la tradición oral moderna portuguesa, tal como dijimos arriba nuestro *corpus* está constituido por textos en distintos idiomas.

3. Diseño y objetivos de Ulisses

Concluimos que necesitaríamos una aplicación que, aparte de permitirnos mantener todas las informaciones extra-textuales de clasificación de los textos, nos posibilitara la corrección de los errores de etiquetaje. Ulisses nace, así, de la necesidad de crear una aplicación integrada que, además de tokenizar y etiquetar morfosintácticamente el *corpus*, nos permitiera:

i. construir una base de datos en donde pudiéramos catalogar y codificar los textos para mantener diversas informaciones extratextuales.

ii. cruzar información/datos del *corpus* y constituir diversos sub-*corpora* (organizados, en nuestro caso, por romance, por tema, por área geográfica o por editor), a partir de los cuales se puede comprobar la existencia o no de léxicos particulares dentro del léxico general.

iii. constituir o bien un *lexicon* para cada nuevo *corpus* o bien un *lexicon* reutilizable por otros investigadores.

iv. crear un léxico/diccionario de formas flexionadas y de lemas presentes en el *corpus* con la respectiva localización; en nuestro caso, con la indicación del romance/versión en la que aparece.

v. editar los textos sin perder las anotaciones ya efectuadas.

vi. eliminar o añadir textos, entradas en el *lexicon*, *tags* y cualesquiera otros campos considerados pertinentes para el análisis del *corpus* en cualquier momento del proceso de etiquetaje y sin necesidad de recurrir a un informático.

vii. escoger autónomamente el conjunto de *tags.*

viii. corregir los errores de etiquetaje detectados.

ix. integrar nuevas herramientas.

En resumen, en la base del proyecto está el deseo de crear una aplicación que nos propocionara, de forma integrada y en un mismo ambiente de trabajo, las herramientas necesarias: un analizador morfosintáctico, un editor de texto, un módulo de *corpus query*, una herramienta de catalogación y atribución de anotaciones a los textos y, no menos importante, una herramienta que nos permitiera rever los textos y corregir errores de etiquetaje.

3.1. Arquitectura de Ulisses

Ulisses se caracteriza, ante todo, por ser un IDE (*Integrated Development Environment*) que proporciona una interfaz 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 cuanto IDE y en su filosofía, se puede comparar al proyecto GATE (General Architecture for Texte Engineering), que se puede consultar en http://gate.ac.uk/, y al proyecto Ellogon, que parte de la propuesta de GATE y se puede consultar en <www.ellogon.org/>[5].

En lo que se refiere a pormenores técnicos, se ha desarrollado 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[6].

3.2. Módulos ya existentes

Ulisses cuenta con distintos módulos: un *tag manager*, un *attributes manager* y un *annotation manager*. Cuenta con un editor de textos, un *corpus manager*, con editores para el lexicon y para los atributos de los textos, de las etiquetas, de los *tokens* y de las anotaciones. Cuenta con una búsqueda simple/avanzada, un *tokenizer* y un *tagger*, ambos basados en un algoritmo muy simple, pudiendo el investigador optar por tokenizar y etiquetar los *tokens* tanto automática como manualmente. Y, por fin, cuenta con un módulo de *Corpus Query* y otro que permite generar un léxico/diccionario.

3.2.1. Tag manager, attributes manager y annotation manager

En Ulisses, el usuario tiene total libertad, siempre que lo desee, de:

i. en el *tag manager*, escoger y definir las *tags* y sus relaciones jerárquicas, los colores a atribuir (o no) a cada *tag* y la terminología lingüística que pretende adoptar, pudiendo añadir, borrar o reorganizar el etiquetario.

ii. en el *attributes manager*, determinar, añadir o borrar los campos de catalogación que considere adecuados al estudio de su *corpus*.

iii. en el *annotation manager*, crear, alterar o borrar todos los campos de anotación que considere relevantes para su *corpus*[7].

[5] No los hemos utilizado porque GATE no ofrece un *tagger* que reconozca el portugués y cuando hemos empezado el análisis de nuestro *corpus* Ellogon se estaba todavía desarrollando. Además, los lenguajes de programación utilizados por GATE y Ellogon fueran un *handicap*, ya que el informático que ha desarrollado el Ulisses no domina ni Java ni TcL, respectivamente.

[6] Nosotros no dominamos ninguno de los lenguajes de programación que nos permitiera la ejecución del proyecto, así que hemos contado con la ayuda de un informático que, gratuitamente, ha desarrollado el Ulisses.

[7] Para el *corpus* romancístico que estudiamos, hemos considerado fundamental crear en el *annotation manager* campos que nos permiten indicar los lemas de cada palabra léxica, en la acepción de Coseriu (1987), y dar cuenta de las formas en final de verso y también de las palabras en castellano/mirandés/gallego.

3.2.2. El editor de textos

En el editor de textos se pueden hacer todas aquellas tareas permitidas y necesarias en un editor de textos: eliminar, añadir y copiar texto o partes de texto; hacer *undo* y *redo*; seleccionar todo un texto o partes de texto; imprimir y previsualizar la impresión. Además permite visualizar y seleccionar *tokens* o seleccionar palabras y convertirlas en *tokens*.

3.2.3. Módulo de búsquedas

Teniendo como principal objetivo el permitir que el proceso de desambiguación y de corrección sea más rápido, el módulo de búsquedas le permite al investigador buscar en el *corpus* palabras *tokens*, lemas, *tags* y otro tipo de anotaciones. El resultado de la búsqueda se presenta en forma de listado con la posibilidad de, a partir de él, acceder directamente al texto en donde figuran las formas buscadas: basta pulsar la respectiva palabra que se pretende ver.

3.2.4. El tokenizer

Como hemos dicho ya, el algoritmo del *tokenizer* es tremendamente simple. El *tokenizer* automático separa los espacios entre palabras, los signos de puntuación y considera también como único *token* todas aquellas formas de la lengua que se separan por guión, con excepción de los pronombres personales adjuntos a verbos, los cuales mantiene como dos *tokens*. El investigador puede, sin embargo, optar por *tokenizar* manualmente un texto o una determinada palabra, corrigiendo posibles errores del proceso automático.

3.2.5. El tagger

Creemos importante volver a recordar aquí que nuestro objetivo no se centra en el desarrollo de un *tagger*, sino en estudiar el léxico del *corpus* con el auxilio de una aplicación informática, aunque para ello necesitemos un *tagger*.

El analizador morfosintáctico automático de Ulisses se basa en un algoritmo probabilístico muy simple: a un token que pueda recibir dos o más etiquetas se le atribuye la etiqueta que ha sido atribuida más veces a lo largo del *corpus* ya etiquetado. Por supuesto que este algoritmo presenta un número significativo de errores, pero se repasa cada texto después de etiquetado para corregir los fallos del etiquetaje.

Con el propósito de que el proceso sea más rápido y para que la interacción con el programa sea lo más práctica posible, el usuario puede acceder a través del teclado a las funcionalidades más frecuentes para repasar todos los *tokens* y todo el etiquetaje, pudiendo manualmente atribuirle a un determinado *token* una nueva etiqueta.

Con respecto al *tagger* y dado que la estructura de Ulisses se encuentra preparada para recibir nuevos módulos, creemos, y deseamos, que muy pronto se le podrá integrar un nuevo analizador que presente un alto nivel de aciertos.

3.2.6. El módulo de Corpus Query

El módulo de *corpus query* permite extraer información cuantitativa de un *corpus* anotado presentándola en forma de listado. La información extraída se puede consultar directamente o

exportar para otros programas para procesamiento y análisis estadísticos[8] (p.ex. en SPSS o en Excel) o para elaboración de documentos (p.ex. en Word).

Este módulo admite que el investigador defina las *queries* que necesita para extraer la información que considera importante para el estudio de su *corpus* y le proporciona un gestor de *queries* con el que crear o mantener un conjunto de *queries* organizado por categoría y descripción para una fácil y rápida reutilización.

Aunque las *queries* se encuentren en lenguaje SQL y por lo tanto no sean muy fáciles para un usuario con pocos conocimientos de informática, el módulo suministra un conjunto de *queries* pre-definidas que le pueden servir como base para la creación de nuevas *queries*, con modificar simplemente algunos de los parámetros.

3.2.7. El módulo diccionario

Partiendo de los criterios definidos para la anotación de los *tokens* del *corpus* o de los atributos de catalogación de los textos, este módulo permite generar un diccionario/vocabulario de lexemas o de lemas (tanto por orden alfabético como por orden decreciente de ocurrencias en el *corpus*) con la indicación de su categoría gramatical y respectivos contextos de ocurrencia en el *corpus*.

La interfaz, muy simple y muy intuitiva, le permite al investigador restringir el contenido del diccionario/vocabulario a las partes del *corpus* que le parezcan más relevantes a través de la especificación de algunos de los criterios basados en los atributos de catalogación de los textos o en las anotaciones atribuidas a los *tokens* del *corpus*[9].

La presentación del diccionario/vocabulario se hace en formato de muy fácil lectura y está pensado para su impresión en papel, pudiendo exportarse en formato xml o importarse en programas de procesamiento de texto.

3.3. Ambiente de trabajo

Una de las particularidades del ambiente de trabajo de Ulisses es su interactividad, que resulta del hecho de que el usuario puede ver los textos del *corpus* organizados por atributos, atributos estos que define en una estructura jerárquica de hasta tres niveles los cuales puede volver a organizar.

También, a fin de que el proceso de desambiguación y de corrección sea más rápido, el programa permite la visualización de las etiquetas atribuidas:

---

[8] Los primeros datos estadísticos de los romances de la tradición oral moderna portuguesa editados entre 1828 y 1960 los hemos presentado en Pires (2005 y 2007b). A su vez, en la tesis doctoral, hemos presentado un estudio estadístico más amplio de las palabras léxicas del *corpus* (sustantivos, adjetivos, verbos, adverbios y cuantificadores) y hemos presentado, además, listas de hápax. Las listas de hápax y los diez gráficos y quince cuadros con datos estadísticos nos han permitido comprobar algunas especificidades del léxico de nuestro *corpus*, las cuales no podemos presentar aquí por problemas de espacio.

[9] En nuestro caso particular, hemos optado por editar todo el léxico (tanto las palabras léxicas como sus lemas) de los romances ya que hasta el presente todavía no se editara el léxico de un *corpus* romancístico. No obstante, partiendo de las anotaciones atribuidas a los *tokens* o partiendo de los atributos de catalogación de los textos podríamos (y lo podremos en todo momento) comparar, entre otros: el léxico de distintos romances (épicos y históricos con carolíngeos o con clásicos o con novelísticos, etc.); el léxico de romances editados por distintos editores o el léxico de romances recopilados en distintas regiones. Por ejemplo, en un pequeño artículo (Pires, 2007a) estudiamos solamente las referencias antroponímicas y toponímicas en los romances editados entre 1828 e 1960, por lo que se pueden adoptar múltiples perspectivas para el estudio del léxico del *corpus*.

i. con la coloración del texto y a través de otras guías visuales (como el subrayado o el negrito de los *tokens* etiquetados).

ii. a través de una lista (paralela al *lexicon*) de los *tokens* del texto con el lema y respectivas etiquetas.

iii. a través de una ventana interactiva en la cual se pueden ver y alterar las etiquetas de un *token* previamente seleccionado en el texto.

De su interactividad se destaca, por último, el hecho de que al usuario se le permita organizar las ventanas de la interfaz con total libertad, pudiendo colocarlas con la disposición que considere más conveniente: minimizándolas, redimensionándolas u ocultándolas.

4. Consideraciones finales

De lo expuesto, nos parece que Ulisses, comparativamente con otras aplicaciones disponibles, se nos presenta como una herramienta bastante poderosa en virtud de que se puede usar tanto en el ámbito de los estudios lingüísticos como en el ámbito de los estudios filológicos, pudiendo ser aplicado a cualesquiera *corpora*.

Sin embargo, en el ámbito de los estudios lingüísticos reconocemos que, en la actualidad, el mayor problema de Ulisses se encuentra en el tiempo empleado en el etiquetaje del *corpus* y en el tiempo empleado en la construcción del *lexicon* a partir del cual se etiquetará automáticamente el *corpus*, siempre y cuando el investigador no opte por importar un *lexicon* ya construido por otros investigadores. Así que nuestro deseo es el de que algún día Ulisses pueda realmente integrar un *tagger* con mejor *performance*.

No obstante, creemos que las ventajas del programa Ulisses las encontramos:

i. en su aplicabilidad a todo tipo de *corpus*, permitiendo mantener todas y cuantas informaciones extratextuales el investigador crea necesarias para el estudio de su *corpus*. Es decir, a través de Ulisses se puede estudiar el léxico de un autor, el léxico de varios autores, el léxico de una o más publicaciones periódicas contemporáneas o de diferentes épocas, el léxico de diferentes géneros periodísticos, etc.

ii. en el hecho de tratarse de una estructura modular a la cual, en todo momento, se le pueden añadir nuevos módulos, como un gestor de concordancias, otro módulo de *POS-Tagging* automático, un *Lemmatizer* automático, un módulo de *Data Mining*, un módulo de análisis estadísticos[10], etc.

iii. en el hecho de que es una aplicación interactiva que permite que cada investigador defina, modifique, añada o elimine los campos de catalogación, el etiquetario y las anotaciones, ajustándolas a la especificidad de su *corpus* y a los objetivos de su trabajo de investigación.

iv. en el hecho de que se puede o bien optar por reutilizar un *lexicon* ya constituido por otro investigador, o bien construir un nuevo *lexicon* a partir de un nuevo *corpus*.

[10] Al Ulisses todavía no se le ha integrado un módulo de estadísticas porque los datos cuantitativos obtenidos en el módulo *Corpus Query* pueden tratarse estadísticamente en otra aplicación. De todos modos, seguimos planteando añadirle pronto un módulo de estadísticas.

v. en el hecho de que permite construir (o bien para todo el *corpus* o bien para una parte el *corpus*) un léxico/diccionario de palabras léxicas, de lemas y/o de palabras gramaticales con la indicación de las respectivas ocurrencias en cada texto del *corpus*.

vi. en el hecho de que el programa posee una interfaz muy intuitiva.

vii. por archivar toda la información en base de datos.

viii. y, por fin, en la sustancial reducción de la dependencia del investigador con relación al informático. Es decir, con esta aplicación el investigador planea y gestiona en la interfaz todo su trabajo, necesitando del informático únicamente para la construcción de nuevos módulos.

Referencias

Afonso S., Bick E., Haber R. y Santos D. (2002). Floresta Sintá(c)tica: um treebank para português. In Gonçalves A. y Correia C. N., editores, *Actas do XVII Encontro da Associação Portuguesa de Linguística*: 533-545.

Araújo T. (1998). Casada em Terras Longínquas no Baixo Alentejo em confronto com outras tradições atlânticas e mediterrânicas. *Arquivo de Beja*, Série III, VII/VIII: 221-227.

Barreiro A., Pereira M. J. y Santos D. (1993). *Critérios e Opções Linguísticas no Desenvolvimento do Palavroso, um Sistema Computacional de Descrição Morfológica do Português*, Grupo de Linguagem Natural do INESC, Relatório INESC nº RT/54-93, <www.linguateca.pt/diana/download/criterios.ps>, pp.1-39.

Correia M. (1996). Terminologia e Lexicografia Computacional. In *Jornada Panllatina de Terminologia*. Barcelona, IULA/Universidade Pompeu Fabra, 83-91.

Coseriu E. (1987). *Gramática, Semántica, Universales*. Madrid, Gredos.

Oksefjell S. y Santos D. (1998). Breve panorâmica dos recursos de português mencionados na Web. In Lima, V. L. S., editora, *Anais do 3º Encontro de Processamento da Língua Portuguesa Escrita e Falada, PROPOR'98*, 38-47.

Pires N. A. (2005). O léxico dos romances carolíngeos da Tradição Oral Moderna portuguesa editados entre 1828 e 1960: uma amostra. In Laranjinha A. S. y Miranda J. C., editores, *Modelo – Actas do V Colóquio da Secção Portuguesa da Associação Hispânica de Literatura Medieval*, 231-242.

Pires N. A. (2007a). Referências antroponímicas e toponímicas em romances de tradição oral moderna portuguesa editados entre 1828 e 1960. *Revista Culturas Populares*, nº 4: < www.culturaspopulares.org/Indices4.html>.

Pires N. A. (2007b). Verbos e tempos verbais nos romances carolíngeos da tradição oral moderna portuguesa, editados entre 1828 e 1960. In López Castro A. y Custa Torre M. L., editores, *Actas do XI Congreso de la Asociación Hispánica de Literatura Medieval*, Vol. 1, 143-152.

Ranchhod E. M. (1999). Dicionários Electrónicos e Análise Lexical Automática. In Marrafa P. y Mota Mª A., editores, *Linguística Computacional – Investigação Fundamental e Aplicações*: 207-233.

Rocha P., Simões A. M. y Almeida J. J. (2002). Cálculo de frequências para entradas de dicionários através do uso conjunto de analisadores morfológicos, taggers e corpora. In Gonçalves A. y Correia C. N., editores, *Actas do XVII Encontro da Associação Portuguesa de Linguística*, 407-418.