De la especificidad de un *corpus* romancístico a la creación de una nueva aplicación: ULISSES

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

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 construir 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

Abstract: On 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'll describe the Ulisses application, the IDE which was specifically developed for the annotation and analysis of the *corpus*.
Keywords: Ballads; lexicon; annotation; *corpus*; Ulisses

1 Introducción

En el ámbito de nuestra tesis de doctorado nos hemos propuesto estudiar el léxico del romancero de la tradición oral moderna portuguesa editado entre 1828 e 1960[1], o sea, un *corpus* constituido por 1721 textos y que contiene versiones de cerca de cien romances distintos recopilados en diferentes regiones y editados por distintos editores a lo largo de 132 años.

Nuestro trabajo de investigación tiene como objetivo principal el estudio del léxico y la construcción de un diccionario de formas flexionadas, con la indicación del romance/versión en el que aparece y presupone, además, la formación de diversos sub-*corpora* a partir del *corpus* principal con el objetivo de verificar si, además del léxico general, se puede hallar un léxico particular de cada romance, de cada región, de cada tema/asunto, de cada informante e incluso de cada editor.

Para el análisis del léxico hemos necesitado anotar todo el *corpus*, ya que nuestros objetivos contemplan el estudio de las clases de palabras más frecuentes (Nombres, Adjetivos, Verbos, Adverbios y Cuantificadores, en particular los numerales), el estudio de los tiempos verbales y demás flexión nominal.

Por lo tanto, 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 construir 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 desarrollado específicamente para la anotación y análisis del *corpus*.

[1] Por cuestiones científicas y metodológicas, en nuestro *corpus* se incluyen únicamente las versiones de los romances tradicionales de asunto profano y de romances devotos tradicionales. Se excluyen, por lo tanto, todas las versiones de romances religiosos, las versiones de romances vulgares de asunto profano y las versiones de romances devotos vulgares publicadas entre 1828 y 1960.

1.1 Especificidad del corpus romancístico

Las singularidades de nuestro *corpus* se hallan en las características específicas del género literario romance: texto poético con versos mayoritariamente monorrimos, con rima asonante. Por lo general, y con el fin último de mantener la rima, la estructura sintáctica de los textos presenta muchas inversiones y no es raro que se encuentren en un mismo verso o en todo un texto cambios en los tiempos verbales que, en realidad, no son permitidos por las reglas de la lengua.

Además, y como hemos referido arriba, en nuestro *corpus* se encuentran versiones de cerca de cien romances distintos recogidos en diferentes regiones y editados por distintos editores a lo largo de 132 años. De este modo, y dado que "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 surgen recitados en castellano, en gallego y en portugués, recitados en castellano/portugués, en gallego/portugués o incluso en castellano/gallego/portugués. Por otra parte, los textos oriundos del municipio de Miranda do Douro y del área de influencia del mirandés/leonés nos surgen recitados en mirandés, en mirandés/portugués, en mirandés/castellano y en mirandés/castellano/portugués.

1.2 Anotación del *corpus*

Los rasgos que pretendíamos anotar en nuestro *corpus* se relacionan directamente con el estudio del léxico general y de los léxicos particulares (si los hay), con la especificidad del género literario que estudiamos y con el hecho de que trabajamos con textos bilingües y trilingües. De este modo 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 esta o no contaminado, un código de versión, un código de tema/asunto y un código de romance.
ii. una herramienta de *full tagging*.
iii. atribuir una anotación a cada *token* en fin de verso.
iv. 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[2].

2 Búsqueda de herramientas

Durante cerca de un año, hemos buscado en la *World Wide Web* soluciones informáticas que nos permitieran cumplir los objetivos definidos para nuestra investigación. Nos hemos dado cuenta de que en todo el mundo existe una gran cantidad de proyectos de PLN iniciados aunque no todos terminados. Muchos más iniciados y terminados para el inglés que para las demás lenguas románicas. Muchos más proyectos concluidos para el Portugués del Brasil (PB) que para el Portugués Europeo (PE). Pero, también nos hemos dado cuenta de que es más sencillo encontrar referencias a los proyectos que poder probarlos y utilizarlos realmente.

Aunque el número de proyectos y de trabajos en el ámbito del PLN[3] para el portugués europeo (PE) hayan aumentado significativamente en los últimos 10 años, siguen siendo pocos los analizadores morfológicos o morfosintácticos disponibles para el PE: EMS (http://natura.di.uminho.pt/~jj/pln/pln.html)[4]; LX-Suite (http://lxsuite.di.fc.ul.pt/) y VISL (http://visl.sdu.dk/). Además, las

[2] En este caso están preposiciones como *de* y *para* (portugués/gallego/castellano/mirandés) o nombres como *Roma*, *pulso*, *barba* y *espada* (portugués/gallego/castellano), entre muchos otros casos.

[3] Para una información más detallada sobre los proyectos que se vienen desarrollando en el ámbito del PLN del PE, consultar Oksefjell y Santos (1998) y los sites <http://acdc.linguateca.pt/treebank/ > y <www.linguateca.pt/>.

[4] Integrado en el proyecto Jspell, que destacamos por ser un proyecto pionero y que además es el único totalmente en *open source* (Rocha, Simões e Almeida, 2002).

herramientas de anotación y análisis que encontramos singuen siendo dispares y su integración difícil.

2.1 Aplicabilidad de los analizadores morfológicos al *corpus* romancístico

Si la anotación de un *corpus* compuesto por textos periodísticos no es, desde luego, una tarea sencilla, anotar un *corpus* romancístico se revela una tarea mucho más compleja por tratarse de texto poético con estructuras sintácticas específicas. Y se vuelve aún más difícil cuando no hay un sólo analizador morfosintáctico entrenado con un *corpus* de las características del nuestro.

Antes de optar por el diseño de una nueva aplicación, aceptando todos los riesgos que tal decisión conlleva, hemos probado con un texto de muestra cada uno de los tres analizadores[5] disponibles para el PE y hemos constatado que VISL fue el único *tagger* con aciertos de 97% (97,3%). Sin embargo, se nos han planteado algunas cuestiones: si bien para estudios estadísticos del léxico un 3% de errores de etiquetaje no resulta significativo, en nuestro caso específico puede implicar un número importante de errores en el diccionario.

En esta fase de trabajo tampoco las hipótesis eran muchas. Podríamos haber optado por entrenar un nuevo *tagger*, pero para su entrenamiento necesitaríamos un *corpus* con las características del nuestro y, en realidad, nunca lo hizo nadie. Y para que nos fueran útiles nuestros textos en el entrenamiento del nuevo *tagger* tendríamos que anotar manualmente gran parte de los 1721 textos del *corpus*.

Podríamos haber adoptado el etiquetaje del analizador VISL, que nos obligaría a repasar todo el output para corregir errores. Pero, no se encuentra disponible ni un ejecutable ni el código fuente de esta aplicación para integrar en Ulisses; y aunque existe una demo *online*, es limitada.

Podríamos haber cambiado los objetivos de nuestro trabajo de investigación. Sin embargo, creíamos posible crear una aplicación que nos permitiera utilizar, de forma integrada y en un mismo ambiente de trabajo, las distintas herramientas que necesitábamos: un analizador morfosintáctico, un editor de texto, un programa de estadística, una herramienta con la que catalogar y atribuir otro tipo de anotaciones a los textos y, no menos importante, una herramienta que nos posibilitara repasar los textos para corregir errores.

3 Diseño y objetivos del Ulisses

Ulisses nace, así, de la necesidad de crear una aplicación integrada que, además de permitir 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[6].
iv. crear un léxico/diccionario de formas flexionadas y de lemas presentes en el *corpus* con la respectiva localización; en nuestro caso, con la

[5] Para poder evaluar el grado de acierto de los analizadores, hemos solicitado a cada uno de ellos el etiquetaje de un mismo texto (una versión totalmente en portugués del romance *Conde Claros en Hábito de Fraile*), contabilizando después, en el *output*, los errores de etiquetaje.

[6] Al posibilitar la importación y la exportación de datos en formato .xml, pretendemos que otros investigadores, si lo desean, puedan reutilizar el *lexicon* ya construido como base del proceso de etiquetaje de nuevos *corpora* o con objetivos investigadores. En el caso de nuestro *lexicon*, creemos que será útil para futuras investigaciones en el ámbito del romancero o en el ámbito del estudio de otros géneros da literatura tradicional.

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 cualesquier 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 herramientas desarrolladas específicamente para determinada investigación

Por fin, en la base del proyecto está el deseo de crear una aplicación que posibilitara que otras personas no tengan que pasar más tiempo en la búsqueda de herramientas y en su adaptación e integración que en hacer un trabajo más útil: su propia investigación.

3.1 Arquitectura del 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/>[7].

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[8].

Actualmente, posee dos versiones con distinta estructura, dado que, inmediatamente después de los primeros tests[9], nos hemos dado cuenta de que serían necesarias profundas alteraciones en la estructura del programa: de una estructura rígida en la que los campos y criterios de anotación y de catalogación estaban predefinidos en el código fuente hemos pasado a una estructura totalmente flexible e interactiva en la que el usuario define todos los campos y criterios de catalogación que necesita.

[7] 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 desarrolló el Ulisses no domina ni Java ni TcL, respectivamente.

[8] 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.

[9] Los primeros datos estadísticos resultado de la aplicación de Ulisses a nuestro corpus los presentamos en Pires (2005a) y Pires (2005b).

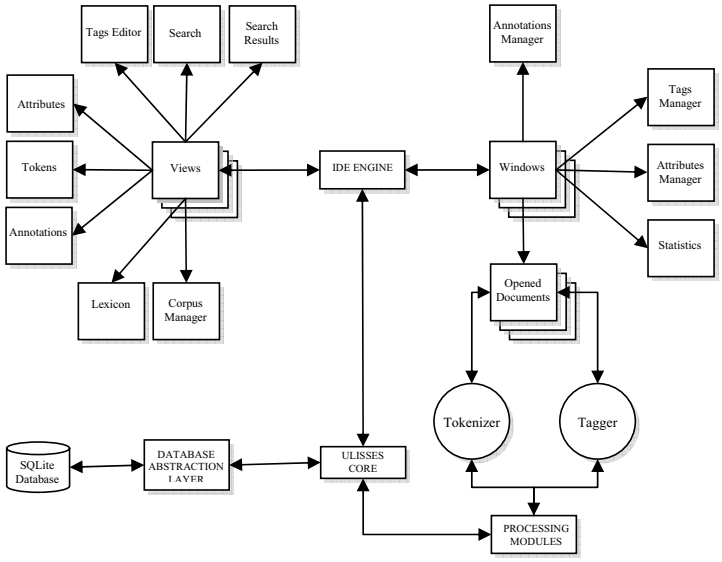

Figura 1 – Diagrama funcional del programa

3.2 Módulos ya existentes

En su segunda versión 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 léxicon y para los atributos de los textos, de las etiquetas, de los *tokens* y de las anotaciones. Y además 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 automáticamente como manualmente.

3.2.1 *Tag manager*, *attributes manager* y *annotation manager*

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

- 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;
- en el *attributes manager*, determinar, añadir o borrar los campos de catalogación que considere adecuados al estudio de su *corpus*;
- en el *annotation manager*, crear, alterar o borrar todos los campos de anotación que considere relevantes para su *corpus*[10].

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 Búsquedas

Teniendo como principal objetivo el permitir que el proceso de desambiguación y de corrección sea más rápido, el módulo

[10] 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.

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.

3.3 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 nivele,s 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:

i. con la coloración del texto y a través de otras guías visuales (como el subrayado o el negro 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.

3.4 Nuevas funcionalidades

La actual versión de Ulisses contará, muy pronto, con:

1 – un módulo de análisis estadístico del *corpus*, incorporando el cálculo de medias, desviaciones, *chi-square*, frecuencias absolutas, frecuencias relativas y otros;

2 – 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.

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 cualesquier *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 acrecentar nuevos módulos, como un gestor de concordancias, otro módulo de *POS-Tagging* automático, un *Lemmatizer* automático, un módulo de *Corpus Query Processor*, un módulo de *Data Mining*, un módulo de análisis estadísticos más complejos, 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*.

v. en el hecho de que permite construir un léxico/diccionario de palabras léxicas 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 el *interface* todo su trabajo, necesitando del informático únicamente para la construcción de nuevos módulos.

Bibliografía

Afonso, Susana, Bick, Eckhard, Haber, Renato e Santos, Diana (2002): "Floresta Sintá(c)tica: um treebank para português", in Gonçalves, Anabela e Correia, Clara Nunes (org.), *Actas do XVII Encontro da Associação Portuguesa de Linguística*, Lisboa, Associação Portuguesa de Linguística, pp. 533-545.

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

Barreiro, Anabela, Pereira, Mª de Jesus y Santos, Diana (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, Margarita (1996): "Terminologia e Lexicografia Computacional", in *Jornada Panllatina de Terminologia*, Barcelona, IULA/Universidade Pompeu Fabra, pp. 83-91.

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

Oksefjell, Signe y Santos, Diana (1998): "Breve panorâmica dos recursos de

português mencionados na Web", in Lima, Vera Lúcia Strube de (ed.), *Anais do 3º Encontro de Processamento da Língua Portuguesa Escrita e Falada, PROPOR'98*, Porto Alegre, pp. 38-47.

Pires, Natália Albino (2005a): "O léxico dos romances carolíngeos da Tradição Oral Moderna portuguesa editados entre 1828 e 1960: uma amostra" in Laranjinha, Ana Sofia y Miranda, José Carlos (eds.), *Modelo – Actas do V Colóquio da Secção Portuguesa da Associação Hispânica de Literatura Medieval*, Porto, Faculdade de Letras da Universidade do Porto, pp. 231-242.

Pires, Natália Albino (2005b): "Verbos e tempos verbais nos romances carolíngeos da tradição oral moderna portuguesa, editados entre 1828 e 1960", in *Actas do XI Congreso de la Asociación Hispánica de Literatura Medieval*, no prelo.

Ranchhod, Elisabete Marques (1999): "Dicionários Electrónicos e Análise Lexical Automática", in Marrafa, Palmira y Mota, Mª Antónia (org.), *Linguística Computacional – Investigação Fundamental e Aplicações*, Lisboa, APL/Edições Colibri, pp. 207-233.

Rocha, Paulo, Simões, Alberto Manuel y Almeida, José João (2002): "Cálculo de frequências para entradas de dicionários através do uso conjunto de analisadores morfológicos, taggers e corpora", in Gonçalves, Anabela e Correia, Clara Nunes (org.), *Actas do XVII Encontro da Associação Portuguesa de Linguística*, Lisboa, Associação Portuguesa de Linguística, pp. 407-418.