

inLéctor: creación de libros electrónicos bilingües interactivos

Antoni Oliver
Universitat Oberta de Catalunya
aoliverg@uoc.edu

Miriam Abuin Castro
Universitat Oberta de Catalunya
mabuinc@uoc.edu

Resumen

En este artículo presentamos el proyecto InLéctor para la creación de libros electrónicos bilingües interactivos. El objetivo del proyecto es desarrollar una serie de aplicaciones para la creación automática de libros electrónicos bilingües. Dichos libros permiten pasar del texto original al traducido con un solo clic y se publican en los formatos de libros electrónicos más habituales: html, epub y mobi. La intención es publicar obras literarias en dominio público (cuyos derechos de autor hayan caducado) con traducciones en dominio público (los derechos de traducción también han caducado). Los programas desarrollados se basan en software libre y se publicarán también bajo una licencia libre. De esta manera las editoriales que lo deseen podrán publicar su catálogo en este formato con una inversión mínima. En un futuro se pretende incluir enlaces al audio correspondiente a la lectura humana de la obra original.

Palabras clave

libros electrónicos, traducción, alineación de textos

Abstract

In this paper we present the project InLéctor, aiming to create interactive bilingual ebooks. The goal of this project is to develop a set of applications for the automatic creation of bilingual ebooks. These books allow to switch from the original text to the translation with a simple click and are published in the more popular ebook formats: html, epub and mobi. We plan to publish literary works in the public domain with translations also in the public domain. All the developed applications are based on free software and will be also published under a free license. Publishing companies will be able to use this software at no cost, which will make possible to publish their catalogue in this format. At this moment we are publishing bilingual ebooks, but in the near future we plan to include links to the audio of the human reading of the original work.

Keywords

ebooks, translations, text alignment

*This work is licensed under a
Creative Commons Attribution 3.0 License*

1 Introducción

El proyecto InLéctor (Oliver, Coll-Florit, y Climent, 2012) pretende fomentar la lectura en versión original, ofreciendo libros bilingües, en texto y audio, en un entorno de lectura interactiva. En este proyecto pretendemos desarrollar una metodología para la creación automática de libros bilingües, utilizando software libre y publicando con licencia libre los programas desarrollados. Además, pondremos a la disposición del público gratuitamente una serie de obras literarias bilingües. Con esta iniciativa esperamos fomentar la publicación de obras literarias bilingües, en dos escenarios:

- Mediante la creación de una comunidad de usuarios que creen nuevas obras literarias en este formato. Las obras literarias que se creen deberán estar en dominio público, tanto el original como su traducción.
- Ofreciendo el software a las empresas editoriales para que puedan publicar su fondo bibliográfico en este formato con una inversión mínima.

Las obras se publicarán en los formatos más extendidos (html, epub y mobi) de manera que se puedan visualizar correctamente en los dispositivos de lectura más habituales

2 Funcionalidades

2.1 Texto bilingüe

Las obras se ofrecerán en su lengua original y en su traducción a otra lengua. El original y la traducción estarán paralelizados a nivel de oración. Esto permitirá un cambio rápido de la versión original a la traducida con un simple clic. De esta manera, si el lector quiere consultar la traducción de una determinada oración, haciendo clic sobre la oración original se visualizará la oración traducida en su contexto, es decir, dentro de la obra traducida. De esta manera, el lector

podrá continuar leyendo la traducción y pasar en cualquier momento al texto original si lo desea.

Si en algún caso no se ha obtenido una alineación de una determinada oración, haciendo clic sobre esta se irá a la oración traducida anterior más cercana. De esta manera nos aseguramos de que todas las oraciones tengan un enlace y el usuario apenas si notará un pequeño desplazamiento del texto.

Para el formato mobi, debido a restricciones de visualización de los enlaces, la alineación se hará a nivel de párrafo y mediante un asterisco entre corchetes ([*]) antes de cada párrafo.

2.2 Audio de la lectura humana del original

Está previsto enlazar el texto original con el audio correspondiente a la lectura humana de la obra en la lengua original. Este aspecto puede ser de gran interés para la mejora de la comprensión oral y la pronunciación. En este caso, la alineación entre el texto y el audio no se realizará a nivel de oración, si no por unidades más amplias aún por determinar (párrafos, conjunto de párrafos o capítulos).

2.3 Glosario interactivo

Antes de iniciar la lectura de un capítulo o fragmento de una obra, el usuario podrá especificar su nivel de lengua lo que le proporcionará un glosario de las palabras más difíciles del texto. El objetivo es que el usuario aprenda el significado de estas palabras más complicadas a priori para así poder disfrutar de una lectura más ágil y con menos interrupciones. Este glosario interactivo se podrá descargar desde la web del proyecto indicando la obra o capítulo y el nivel de lengua del usuario.

Para generar los diccionarios bilingües se utilizarán fuentes libres como por ejemplo Wiktionary¹, WordNets libres (Bond y Paik, 2012) y los diccionarios de transferencia de Apertium (Forcada, Tyers, y Ramírez, 2009).

También se generarán diccionarios bilingües y monolingües libres de manera que se puedan utilizar como diccionario por defecto en los diferentes dispositivos. De esta manera el usuario podrá consultar el significado o la definición de una palabra en cualquier momento. Si el dispositivo ya cuenta con diccionarios bilingües instalados para el par de lenguas en uso, estos podrán utilizarse con las obras de nuestro proyecto, ya que utilizan formatos estándar.

¹<http://www.wiktionary.org/>

2.4 Lectura ampliada

Por *lectura ampliada* entendemos la aproximación a un texto con la ayuda de información adicional que estará disponible al lector mediante un solo clic. La información adicional puede ser de tipo enciclopédico, visual o sonoro. Algunos ejemplos pueden ser la información sobre un lugar geográfico o persona que aparece en el texto; la visualización de una obra de arte o paisaje y la reproducción de un fichero de audio musical relacionada con la obra.

Pongamos un ejemplo hipotético: el protagonista de la obra que estamos leyendo entra en un museo y observa una determinada obra de arte mientras que en el hilo musical se reproduce una determinada sinfonía. El sistema de lectura ampliada enlazaría directamente con la página web del museo, una imagen y explicación de la obra de arte y daría la posibilidad de escuchar la misma sinfonía mientras leemos el pasaje.

2.5 Interacción entre los usuarios

La web del proyecto incluirá una serie de funcionalidades que permitan la interacción de los usuarios. El sistema permitirá compartir comentarios sobre la obra o dudas sobre un determinado fragmento. Esta interacción se llevará a cabo mediante del uso de redes sociales, como Facebook o Twitter.

3 Obtención de obras literarias

Las obras que se publiquen dentro de este proyecto serán únicamente aquellas que tengan los derechos de autor y de traducción libres, es decir, que estén en dominio público. De esta manera, podemos garantizar que la distribución de las obras sea totalmente legal. Así, las obras literarias en versión original y las traducciones se extraerán principalmente de Wikisource² y del Proyecto Gutenberg³. Aunque la caducidad de los derechos de autor depende de cada legislación, de manera general se puede considerar que una obra es de dominio público si han pasado más de setenta años desde la muerte de su autor⁴. Concretamente en España la legislación dicta que las obras quedan en dominio público ochenta años después de la muerte del autor si éste murió antes del 7 de diciembre de 1987, y setenta años después de su muerte si este falleció después de

²<http://wikisource.org/>

³<http://www.gutenberg.org/>

⁴Se puede encontrar información muy detallada en http://en.wikisource.org/wiki/Help:Public_domain

la mencionada fecha.

En cuanto a los audios, serán en su mayoría provenientes de LibriVox⁵, proyecto en el que un gran número de voluntarios leen capítulos de libros que están bajo dominio público, y donde se publican también bajo dominio público los ficheros de audio.

4 Software utilizado

4.1 Software general

Para la creación de los libros electrónicos se necesita contar con una serie de software general que está disponible con licencia libre. Entre ellos es necesario contar con un buen editor de textos. Existen muchísimas opciones y dependerán del sistema operativo que se utilice. Algunos ejemplos son Jedit⁶ (multiplataforma) o notepad++⁷ (para Windows). Lo importante es que el editor elegido cuente con un buen soporte para la creación de macros. También puede ser útil la utilización de editores específicos para XML, como por ejemplo XMLCopyEditor⁸, o herramientas de validación de XML como *xmllint*.

4.2 Entorno de programación

Los programas están desarrollándose en Python⁹ en combinación con el *Natural Language Toolkit*¹⁰ (NLTK) (Loper y Bird, 2002). NLTK es un conjunto de bibliotecas y programas para el Procesamiento del Lenguaje Natural. Mediante NLTK se pueden programar fácilmente las tareas más habituales de procesamiento del lenguaje. En nuestro proyecto lo estamos utilizando para la segmentación de las obras en oraciones.

4.3 Alineación automática

La alineación de los textos correspondientes a la obra original y traducido la realizamos con el alineador automático Hunaling (Varga et al., 2007). Este alineador nos permite obtener las relaciones entre la frase original y la frase traducida. No todas las frases originales obtienen alineaciones con una o más frases traducidas. Esto se puede deber a dos motivos: o bien la frase original no tiene una traducción o bien el programa no ha podido obtener una alineación válida.

Los resultados de la alineación automática mejoran notablemente si en lugar de alinear los textos en sí, alineamos una versión lematizada de los textos. De esta manera, el número de formas se reduce notablemente ya que todas las variantes morfológicas se reducen a un mismo lema. Si además proporcionamos diccionarios bilingües al programa de alineación, los resultados también pueden mejorar considerablemente.

Aunque no todas las oraciones originales obtengan una alineación válida, todas ellas tendrán un enlace a una oración correspondiente a la traducción. Este enlace será la oración traducida en caso de obtener alineación, o bien la oración traducida inmediatamente anterior que haya obtenido una alineación válida.

4.4 Creación de libros electrónicos

Existen una gran variedad de software libre para la creación de libros electrónicos. Hasta el momento hemos utilizado Calibre¹¹, ya que permite la creación de libros en formato epub y mobi, tanto desde una interfaz gráfica de usuario, como desde un terminal, lo que permite automatizar aún más el proceso.

Existen otras opciones para la creación de libros electrónicos, especialmente si se pretende utilizar el formato epub (que es el estándar libre). Una de ellas es Sigil¹², un editor gráfico para la creación de libros en formato epub. También debe destacarse python-epub-builder¹³, un paquete de Python que permite desarrollar programas que automaticen totalmente la creación de libros en formato epub.

4.5 Procesado lingüístico

El procesado lingüístico necesario para la creación de libros paralelos se concreta en dos tareas: la segmentación en oraciones y la lematización.

La segmentación en oraciones es necesaria como paso previo a la alineación. Este paso es decisivo para lograr una buena alineación, y de ningún modo debe considerarse una tarea trivial. Hasta el momento estamos utilizando un segmentador genérico proporcionado por el NLTK. Probablemente, en futuras versiones del sistema, se tendrá que profundizar en este aspecto.

Para el lematizado de los textos se han utilizado etiquetadores morfosintácticos que nos proporcionan, para cada palabra, su lema y una etiqueta morfosintáctica. De esta información nos

⁵<http://librivox.org/>

⁶<http://www.jedit.org/>

⁷<http://notepad-plus-plus.org/>

⁸<http://xml-copy-editor.sourceforge.net/>

⁹<http://www.python.org>

¹⁰<http://nltk.org/>

¹¹<http://calibre-ebook.com/>

¹²<http://code.google.com/p/sigil/>

¹³<http://code.google.com/p/python-epub-builder/>

quedaremos con el lema para crear una versión lematizada de los textos. En las primeras obras estamos utilizando Treetagger (Schmid, 1994), pero se prevee hacer pruebas con Freeling (Carreras et al., 2004) para evaluar si los resultados son más satisfactorios. A medida que aumente el número de lenguas tratadas tendremos que incorporar progresivamente nuevos etiquetadores. No obstante, hay que tener en cuenta que el paso de lematización es opcional.

5 Proceso

5.1 Obtención de las obras

El primer paso que llevaremos a cabo será la obtención de las obras, tanto el original como la traducción. En este paso obtenemos un archivo de texto correspondiente a la obra original y otro archivo correspondiente a la obra traducida. Por el momento simplemente descargamos las obras desde las webs correspondientes. Para agilizar el proceso, en un futuro utilizaremos los *dumps* en XML de Wikisource, que se pueden descargar libremente. Existe un *dump* específico para cada lengua. A partir de estos ficheros podremos obtener las obras originales, y saber a qué lenguas están traducidas. A partir de la información sobre las traducciones podremos acceder a los textos traducidos a partir de los *dumps* correspondientes a cada lengua.

A continuación podemos observar un fragmento de obra en el idioma original:

```
A SCANDAL IN BOHEMIA
I.
To Sherlock Holmes she is always THE woman. I have
seldom heard him mention her under any other name.
In his eyes she eclipses and predominates the whole
of her sex. It was not that he felt any emotion akin
to love for Irene Adler...
```

y el mismo fragmento correspondiente a la traducción:

```
ESCÁNDALO EN BOHEMIA
1.
Ella es siempre, para Sherlock Holmes, la mujer. Rara
vez le he oído hablar de ella aplicándole otro nombre.
A los ojos de Sherlock Holmes, eclipsa y sobrepasa a
todo su sexo. No es que haya sentido por Irene Adler
nada que se parezca al amor...
```

5.2 Transformación de los textos en ficheros docbook

Docbook¹⁴ (Walsh y Muellner, 1999) es un formato estándar basado en XML que nos permite

representar la estructura lógica de un libro. De esta manera se separa totalmente el contenido del formato. La transformación la realizaremos mediante una serie de macros de un editor de texto. Siguiendo el ejemplo anterior, una vez transformados en docbook, los fragmentos tendrían el siguiente aspecto:

```
<chapter>
<title>A SCANDAL IN BOHEMIA</title>
<section>
<title>I.</title>
<para>To Sherlock Holmes she is always THE woman. I
have seldom heard him mention her under any other
name. In his eyes she eclipses and predominates the
whole of her sex. It was not that he felt any emotion
akin to love for Irene Adler...</para>
...

```

y el mismo fragmento correspondiente a la traducción:

```
<chapter>
<title>ESCÁNDALO EN BOHEMIA</title>
<section>
<title>1.</title>
<para>Ella es siempre, para Sherlock Holmes, la mujer.
Rara vez le he oído hablar de ella aplicándole otro
nombre. A los ojos de Sherlock Holmes, eclipsa y
sobrepasa a todo su sexo. No es que haya sentido por
Irene Adler nada que se parezca al amor.</para>

```

Este paso es en realidad opcional, pero preferimos disponer de las obras en este formato ya que existen muchas aplicaciones para transformar documentos docbook en diferentes formatos de salida: html, pdf e incluso epub. El proceso de creación del docbook nos permite revisar a la vez el documento y verificar que la descarga se haya realizado correctamente.

5.3 Segmentación

El siguiente paso consiste en segmentar los textos y transformarlos en un formato de texto que sea adecuado para el alineador automático de textos. Por el momento utilizamos un segmentador genérico del paquete NLTK que nos proporciona buenos resultados. El fichero de salida contiene una oración por línea y adicionalmente, se marcan los párrafos con la marca $\langle p \rangle$. Los fragmentos anteriores tendrían el siguiente aspecto:

```
A SCANDAL IN BOHEMIA
<p>
I.
<p>
To Sherlock Holmes she is always THE woman.
I have seldom heard him mention her under any other
name.
In his eyes she eclipses and predominates the whole
of her sex.
It was not that he felt any emotion akin to love for
Irene Adler.
```

¹⁴<http://docbook.org/>

y el mismo fragmento correspondiente a la traducción:

ESCÁNDALO EN BOHEMIA

```
1.
<p>
Ella es siempre, para Sherlock Holmes, la mujer.
Rara vez le he oído hablar de ella aplicándole otro
nombre.
A los ojos de Sherlock Holmes, eclipsa y sobrepasa a
todo su sexo.
No es que haya sentido por Irene Adler nada que se
parezca al amor.
```

5.4 Lematización

Los textos obtenidos en el paso anterior ya se podrían alinear sin problemas. Ahora bien, para mejorar los resultados de la alineación es aconsejable lematizar los textos. Para ello se pueden utilizar diversas herramientas, por ejemplo Tree-Tagger o Freeling. Siguiendo con el ejemplo anterior, obtendríamos el siguiente resultado:

A SCANDAL IN BOHEMIA

```
<p>
I.
<p>
to Sherlock Holmes she be always the woman .
I have seldom hear him mention her under any other
name .
in his eye she eclipse and predominate the whole
of her sex .
it be not in he feel any emotion akin to love for
Irene Adler .
```

y el mismo fragmento correspondiente a la traducción:

ESCÁNDALO EN BOHEMIA

```
1.
<p>
ella ser siempre para Sherlock Holmes el mujer .
raro vez él haber oír hablar de él aplicar otro
nombre .
a el ojo de Sherlock Holmes eclipsa y sobrepasar
a todo suyo sexo .
no ser que haber sentido por Irene Adler nada que
se parecer al amor .
```

5.5 Alineación

Una vez descargado e instalado Hunalign en nuestro ordenador, ya podremos ejecutar el proceso de alineación mediante una simple orden:

```
hunapertium -utf8 -realign diccionario.dic original.txt
traduccion.txt > alineacion.txt
```

El parámetro *realign* es opcional pero puede mejorar los resultados. Es imprescindible indicar un archivo de diccionario. Los archivos de diccionario tienen la siguiente forma (si es por ejemplo un diccionario para una alineación de inglés a castellano)

```
científico @ scientific
escultórico @ sculptural
estacional @ seasonal
```

Como se puede observar en primer lugar aparecen las palabras en la lengua de llegada. En el caso de no disponer de un diccionario para el par de lenguas de trabajo, se debe indicar igualmente un diccionario que puede ser un fichero vacío.

Los ficheros *original.txt* y *traducción.txt* son los ficheros de texto correspondientes al original y la traducción, que pueden ser simplemente segmentados o bien segmentados y lematizados.

El fichero de alineación tiene el siguiente aspecto:

```
4 0 0
5 1 -0.3
6 1 0
7 2 -0.3
8 2 0
9 2 1.45135
10 3 1.21031
11 4 0.906857
12 5 1.008
```

Se indica la relación entre el número de segmento del fichero original y del fichero traducido y un *score* que indica la calidad de la alineación. Para recuperar un fichero de alineación formado por las oraciones podemos utilizar el *script* llamado *ladder2text.py* que se distribuye con Hunalign.

```
python ladder2text.py alineacion.txt original.txt
traduccion.txt > alineacion_texto.txt
```

Como el orden de los segmentos en la versión lematizada y sin lematizar del original y traducción segmentados son los mismos, si hemos realizado la alineación a partir de los textos lematizados podemos obtener ahora el fichero de alineación con las oraciones sin lematizar indicando en este paso el nombre de los archivos sin lematizar. Al final obtenemos un fichero que relaciona los segmentos originales y traducidos.

```
1.56762 A SCANDAL IN BOHEMIA ESCÁNDALO EN BOHEMIA
```

```
1.89878 I 1
```

```
1.45135 To Sherlock Holmes she is always THE woman.
Ella es siempre, para Sherlock Holmes, la mujer.
```

```
1.21031 I have seldom heard him mention her under any
other name. Rara vez le he oído hablar de ella aplicándole
otro nombre.
```

```
0.906857 In his eyes she eclipses and predominates the
whole of her sex. A los ojos de Sherlock Holmes, eclipsa
y sobrepasa a todo su sexo.
```

```
1.008 It was not that he felt any emotion akin to love
for Irene Adler. No es que haya sentido por Irene Adler
nada que se parezca al amor.
```

5.6 Creación del html bilingüe

Hemos creado un programa en Python que a partir de los documentos en formato docbook correspondiente al original y la traducción y del fichero de alineación genera un fichero html bilingüe. En este fichero aparecen los segmentos originales y traducidos enlazados. El fichero tiene el siguiente aspecto:

```
<p><a name="s-6"/><a href="#t-2">To Sherlock Holmes she
is always THE woman. </a><a name="s-7"/><a href="#t-3">
I have seldom heard him mention her under any other
name.</a>
<a name="s-8"/><a href="#t-4">In his eyes she eclipses
and predominates the whole of her sex. </a>
...
<p><a name="t-2"/><a href="#s-6">Ella es siempre, para
Sherlock Holmes, la mujer. </a> <a name="t-3"/>
<a href="#s-7">Rara vez le he oído hablar de ella
aplicándole otro nombre.</a><a name="t-4"/>
<a href="#s-8">A los ojos de Sherlock Holmes, eclipsa y
sobrepasa a todo su sexo. </a><a name="t-5"/>
<a href="#s-9">No es que haya sentido por Irene Adler
nada que se parezca al amor. </a></p>
```

En el proceso de alineación es posible que algunos segmentos queden sin alinear. En el caso de no encontrar una alineación válida para un segmento se enlaza con el segmento alineado inmediatamente anterior. De esta manera conseguimos que todos los segmentos estén enlazados entre sí.

En el caso de desear crear un libro en formato mobi, se crea un html bilingüe especial, en el que los enlaces se realizan entre párrafos y mediante una marca formada por un asterisco entre corchetes ([*]). Esto es debido a que tenemos un menor control sobre la salida en mobi y no podemos evitar que los enlaces aparezcan subrayados.

5.7 Creación de los libros electrónicos

Mediante la herramienta Calibre podemos transformar los archivos html bilingües en epub y mobi. También podremos añadir ciertos metadatos y una portada. Todas estas operaciones se pueden realizar desde la interfaz gráfica de usuario, pero puede resultar más cómodo hacerlo en el terminal:

Transformación a epub:

```
ebook-convert sherlock_holmes.html sherlock_holmes.epub
--input-encoding=utf-8 --change-justification=justify
--insert-blank-line --chapter=h1 --chapter-mark=pagebreak
--title="The Adventures of Sherlock Holmes"
--authors="Arthur Conan Doyle" --author-sort=Doyle
--publisher=InLéctor --language=en --cover=portada.jpg
```

Transformación a mobi (utilizaremos el html bilingüe especial alineado por párrafos):

```
ebook-convert sherlock_holmes.html sherlock_holmes.mobi
--input-encoding=utf-8 --change-justification=justify
--insert-blank-line --chapter=h1 --chapter-mark=pagebreak
--title="The Adventures of Sherlock Holmes"
--authors="Arthur Conan Doyle" --author-sort=Doyle
--publisher=InLéctor --language=en --cover=portada.jpg
```

6 Obras disponibles

Las lenguas de trabajo iniciales de este proyecto serán el inglés, francés y ruso al castellano o catalán dependiendo de la disponibilidad de las traducciones. Las primeras obras publicadas son:

- The Adventures of Sherlock Holmes (Sir Arthur Conan Doyle) (inglés-castellano)
- Sense and Sensibility (Jane Auste) (inglés-castellano)
- Les Trois Mousquetaires (Alexandre Dumas) (francés-castellano)
- Ирок (El jugador) de Fiódor Dostoyevski (ruso-castellano)

En un futuro próximo se trabajará con más lenguas. Para cada obra literaria seleccionada para su publicación se editarán libros bilingües en todas las lenguas cuya traducción esté disponible en Wikisource.

7 Conclusiones y trabajo futuro

En este artículo hemos presentado una metodología automática para la creación de libros electrónicos bilingües, desarrollada en el marco del proyecto InLéctor de la Universitat Oberta de Catalunya. Actualmente el mercado del libro se encuentra inmerso en un cambio de paradigma y el paso de formato papel a formato digital. El avance en este cambio es lento, al menos en nuestro país, ya que en pocos casos la edición digital comporta una mejora substancial para usuario final, ni en precio, ni en funcionalidades. Creemos que dotar al libro electrónico con las funcionalidades previstas en este proyecto puede suponer un impulso para el libro electrónico en nuestro país.

En nuestro proyecto tratamos únicamente con obras en dominio público, tanto en lo que hace referencia al original como a la traducción. Las editoriales que dispongan de los derechos de autor y de traducción de una obra se pueden beneficiar de nuestras propuestas y publicar las obras de su catálogo en este formato. El libro digital se asemejaría a una película en DVD, al hacer posible que el usuario pueda escoger la lengua

y los subtítulos y adaptar la visualización a sus preferencias.

Actualmente el proyecto no cuenta con financiación específica por lo que el avance es lento. Nuestra intención es obtener financiación externa para mejorar las herramientas de creación de libros paralelos. El equipo investigador está abierto a colaboraciones externas ya sea en la mejora de las herramientas como en la creación de nuevas obras.

Los siguientes pasos que llevaremos a cabo serán en las siguientes direcciones:

- Automatizar la obtención de las obras originales y traducidas mediante los *dump xml* de Wikisource
- Mejorar el algoritmo de creación de libros electrónicos
- Añadir enlaces al audio correspondiente a la lectura humana del original

Las nuevas especificaciones del formato epub3 (Garrish, 2012) facilitarán enormemente la integración de todas estas funcionalidades.

Una vez desarrolladas las funcionalidades previstas y creadas un número suficiente de obras se pretende fomentar el uso de estas obras en aulas de enseñanza de idiomas. En las últimas etapas del proyecto llevarán a cabo experimentos psicolingüísticos para evaluar las mejoras en la adquisición de léxico y estructuras gramaticales por parte de los estudiantes de idiomas.

Bibliografía

- Bond, Francis y Kyonghee Paik. 2012. A survey of wordnets and their licenses. En *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, páginas 64–71, Matsue, Japan.
- Carreras, X., I. Chao, L. Padró, y M. Padró. 2004. Freeling: An open-source suite of language analyzers. En *Proceedings of the 4th LREC*, volumen 4.
- Forcada, M. L, F. M Tyers, y G. Ramírez. 2009. The apertium machine translation platform: five years on. En *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, página 3–10.
- Garrish, Matt. 2012. *Accessible EPUB 3*. O'Reilly Media.
- Loper, E. y S. Bird. 2002. NLTK: the natural language toolkit. En *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, página 63–70.
- Oliver, A., M. Coll-Florit, y S. Climent. 2012. Inléctor: Sistema de lectura bilingüe interactiva. *Procesamiento del Lenguaje Natural*, 49:279–286, September.
- Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. En *Proceedings of international conference on new methods in language processing*, volumen 12, página 44–49.
- Varga, D., P. Halácsy, A. Kornai, V. Nagy, L. Németh, y V. Trón. 2007. Parallel corpora for medium density languages. En *Proceedings of the RANLP 2005*, páginas 590–596.
- Walsh, N. y L. Muellner. 1999. *DocBook: The definitive guide*, volumen 1. O'Reilly Media.