Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 203-206 recibido 05-05-12 revisado 31-05-12 aceptado 05-06-12

InLéctor: Sistema de lectura bilingüe interactiva*

InLéctor: a system for bilingual interactive reading

Antoni Oliver, Marta Coll-Florit, Salvador Climent
Universitat Oberta de Catalunya
Avda. Tibidabo 39-43 08035 Barcelona
aoliverg,mcollfl,scliment@uoc.edu

Resumen: Este proyecto pretende desarrollar un sistema que genere libros bilingües, con audio e interactivos. El sistema ofrecerá diversos formatos de salida que permitan leer y escuchar los libros en diferentes dispositivos, como libros electrónicos, tabletas y ordenadores. Asimismo, ofrecerá la posibilidad de obtener libros paralelos impresos.
Palabras clave: libro electrónico, aprendizaje de lenguas

Abstract: The aim of this project is the development of a system for the generation of interactive bilingual electronic books with audio support. The system will offer several output formats for reading and listening the books on different devices such as electronic books, tablets and computers. It will also allow printing a parallel bilingual book.
Keywords: electronic book, language learning

1. Introducción

El proyecto InLéctor pretende fomentar la lectura en versión original, ofreciendo libros bilingües, en texto y audio, en un entorno de lectura interactiva. En este proyecto queremos fomentar la lectura en lenguas extranjeras ofreciendo una interactividad que permita conocer a priori las palabras más complicadas de un capítulo; enlaces al párrafo traducido, lo que permitirá al usuario pasar de la versión original a la versión traducida de manera inmediata; y ofrecer en algunas obras el audio correspondiente a una lectura humana.

2. Funcionalidades

2.1. Glosario interactivo

Antes de iniciar la lectura de un capítulo o fragmento de una obra, el usuario podrá especificar su nivel de lengua y obtendrá un glosario de las palabras más difíciles del texto. El objetivo es que el usuario aprenda el significado de estas palabras difíciles para así poder disfrutar de una lectura más ágil y con menos interrupciones.

2.2. Texto bilingüe

Las obras se ofrecerán en su lengua original y en su traducción a otra lengua (para nuestro proyecto el castellano o catalán). El original y la traducción estarán paralelizados a nivel de párrafo. Esto permitirá un cambio rápido de la versión original a la traducida.

2.3. Interacción entre los usuarios

El sistema permitirá compartir comentarios sobre la obra o dudas sobre un determinado fragmento que serán accesibles para todos los usuarios.

2.4. Recursos y herramientas

Las obras que se publiquen bajo este proyecto serán únicamente las que tengan los derechos de autor y de traducción libres, es decir, que estén en dominio público. De esta manera nos aseguramos que la distribución de las obras sea totalmente legal. Así, las obras literarias en versión original y las traducciones se extraerán principalmente de Wikisource[1] y del Proyecto Gutemberg[2].

* Este trabajo se ha llevado a cabo dentro del proyecto Know2 *Language understanding technologies for multilingual domain-oriented information access* (MICINN, TINN2009-14715-C04-04)

[1] http://wikisource.org/
[2] http://www.gutenberg.org/

ISSN 1135-5948

En cuanto a los audios, serán en su mayoría provenientes de la fuente LibriVox[3]. Librivox es un proyecto en el que un gran número de voluntarios leen capítulos de libros que están bajo dominio público, y publican también bajo dominio público los ficheros de audio.

Para generar los diccionarios bilingües se utilizarán fuentes libres como por ejemplo Wiktionary[4], WorNets libres (Bond y Paik, 2012) y los diccionarios de transferencia de Apertium (Forcada, Tyers, y Ramírez, 2009).

En cuanto al procesamiento de los textos, se realizará un etiquetado morfosintáctico con Freeling (Carreras et al., 2004) o Treetager (Schmid, 1994). Para obtener los texto paralelos se utilizará el alineador automático Humaling (Varga et al., 2007) o mAligna (Jassem y Lipski, 2008).

2.5. Formatos

El formato básico de trabajo será el DocBook, con atributos especiales que permitan la relación entre los diferentes párrafos de las dos versiones del texto. A partir de este formato básico se generarán los siguientes formatos de salida: HTML, Mobipocket, Epub y PDF.

2.6. Lenguas de trabajo y obras

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 que está previsto tratar son: The adventures of Sherlock Holmes (Sir Arthur Conan Doyle), Les Trois Mousquetaires (Alexandre Dumas) y Игрок (El jugador) de Fiódor Dostoyevski con sus respectivas traducciones al castellano.

2.7. Público objetivo

Este sistema está pensado para todas aquellas personas con un nivel medio-avanzado de una lengua extranjera que deseen leer obras en versión original. El sistema puede ayudar a mantener un nivel adecuado de la lengua, de una manera amena. Su uso en docencia es interesante, ya que se puede aplicar tanto en el estudio de lenguas extranjeras, como en el estudio de la literatura e incluso en estudios de traducción.

3. Flujo de trabajo

En esta sección expondremos los detalles sobre el flujo de trabajo que estamos siguiendo para la creación de los libros paralelos. Por el momento estamos en la fase de creación de los enlaces entre el párrafo original y el párrafo traducido. Todavía no estamos trabajando la parte de enlaces entre el texto y el audio correspondiente a la locución humana del texto.

3.1. Descarga de la obra

Este primer paso no entraña ninguna dificultad. El objetivo es disponer del original y la traducción en sendos ficheros de texto plano.

```
THE ADVENTURES OF SHERLOCK HOLMES
Arthur Conan Doyle
ADVENTURE I. A SCANDAL IN BOHEMIA
I.
To Sherlock Holmes she is always THE
woman. I have seldom heard him mention
her under any other...
```

3.2. Creación de los ficheros Docbook del original

A partir del fichero de texto correspondiente a la obra en la lengua original se genera un fichero en formato Docbook. Este formato permite crear documentos en un formato independiente de la presentación final. En el formato Docbook se expresa tanto el contenido como la estructura lógica del documento, pero no su formato final. Para crear el fichero en formato Docbook se puede utilizar cualquier editor de textos que tenga un buen soporte de macros (en nuestro caso hemos utilizado JEdit[5]. Mediante la creación de unas pocas macros la conversión de texto plano a Docbook se puede realizar rápidamente.

[3] http://librivox.org/
[4] http://www.wiktionary.org/
[5] http://www.jedit.org/

```
<book>
<title>THE ADVENTURES OF SHERLOCK HOLMES</title>
<chapter>
<title>ADVENTURE I. 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...</para>
...
```

3.3. Creación del Docbook numerado

Mediante un sencillo *script* se numeran los títulos y los párrafos. Esta numeración es la que nos permitirá realizar los enlaces entre la versión original y traducida. Como podemos observar en el siguiente ejemplo, se sigue una numeración independiente para los títulos (ya sean de libro, capítulo o sección) y para los párrafos.

```
<book>
<title xml:id="t1-eng">THE ADVENTURES
OF SHERLOCK HOLMES</title>
<chapter>
<title xml:id="t2-eng">ADVENTURE I.
A SCANDAL IN BOHEMIA</title>
<section>
<title xml:id="t3-eng">I.</title>
<para xml:id="p1-eng">To Sherlock Holmes
she is always THE woman. I have seldom
heard him mention her under any other...</para>
...
```

3.4. Alineación y creación del TMX

Mediante los algoritmos de alineación citados anteriormente se alinean los ficheros de texto correspondientes al original y a la traducción. Aunque la alineación que nos interesa es a nivel de párrafo, la alineación se lleva a cabo a nivel de oración. A partir de la alineación se creará un fichero de memoria de traducción en el formato estándar TMX (*Translation Memory eXchange*).

```
<tu>
<tuv xml:lang="en">
<seg>
To Sherlock Holmes she is always THE woman.
</seg>
</tuv>
<tuv xml:lang="es">
<seg>
Ella es siempre, para Sherlock Holmes, la mujer.
</seg>
</tuv>
</tu>
<tu>
```

3.5. Creación del proyecto de traducción

En este paso creamos un proyecto de traducción con alguna herramienta de traducción asistida. Dado que todos los formatos de fichero son estándar, se puede utilizar cualquier herramienta. En nuestro proyecto utilizamos OmegaT[6] que es una herramienta de software libre. En el proyecto asignamos como documento a traducir el Docbook numerado correspondiente a la obra original y como memoria de traducción el fichero TMX creado en el paso anterior.

3.6. Verificación de la alineación

La verificación de la alineación se ha convertido en una tarea de traducción mediante una herramienta de traducción asistida donde la inmensa mayoría de las oraciones se traducirán directamente mediante la memoria de traducción. El encargado de verificar la alineación sólo tiene que preocuparse de verificar que la propuesta proveniente de la memoria sea la adecuada, lo que ocurrirá en la mayoría de casos. En los casos en que la propuesta no sea correcta o que simplemente no aparezca ninguna propuesta, el encargado podrá consultar el fichero correspondiente a la obra traducida para encontrar la traducción al segmento correspondiente.

3.7. Creación del Docbook numerado correspondiente a la traducción

La creación del Docbook numerado correspondiente a la traducción se reduce a la creación del fichero traducido mediante la herramienta de traducción asistida. Mediante la función de buscar y reemplazar de cualquier editor de textos reemplazaremos las marcas de lengua de la numeración de títulos y párrafos por la marca de lengua correspondiente.

[6] www.omegat.org

```
<book>
<title xml:id="t1-spa">LAS AVENTURAS
DE SHERLOCK HOLMES</title>
<chapter>
<title xml:id="t2-spa">AVENTURA I.
ESCÁNDALO EN BOHEMIA</title>
<section>
<title xml:id="t3-spa">I.</title>
<para xml:id="p1-spa">Ella es siempre,
para Sherlock Holmes, la mujer. Rara
vez le he oído hablar de ella
aplicándole otro ...</para>
...
```

3.8. Creación de los formatos finales

A partir de los ficheros docbook numerado correspondientes al original y a la traducción creamos, mediante un simple *script* el libro en formato html con los enlaces entre los párrafos originales y traducidos.

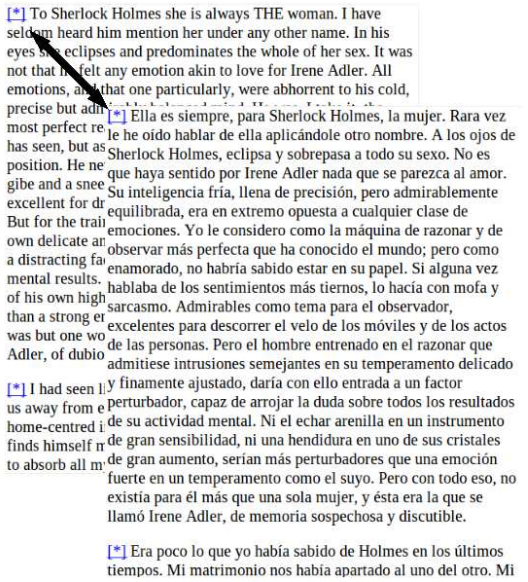
```
<h1><a name="t1-eng"/><a href="#t1-spa">[*]</a>
THE ADVENTURES OF SHERLOCK HOLMES</h1>
<h2><a name="t2-eng"/><a href="#t2-spa">[*]</a>
ADVENTURE I. A SCANDAL IN BOHEMIA</h2>
<h3><a name="t3-eng"/><a href="#t3-spa">[*]</a>
I.</h3>
<p><a name="p2-eng"/><a href="#p2-spa">[*]</a>
To Sherlock Holmes she is always THE woman. I
have seldom heard him mention her under any...
....
....
<h2><a name="t1-spa"/><a href="#t1-eng">[*]</a>
LAS AVENTURAS DE SHERLOCK HOLMES</h2>
<h2><a name="t2-spa"/><a href="#t2-eng">[*]</a>
AVENTURA I. ESCÁNDALO EN BOHEMIA</h2>
<h3><a name="t3-spa"/><a href="#t3-eng">[*]</a>
I.</h3>
<p><a name="p2-spa"/a><a href="#p2-eng">[*]</a>
Ella es siempre, para Sherlock Holmes, la mujer.
Rara vez le he oído hablar de ella aplicándole..
```

A partir de este html con enlaces se utiliza el programa Calibre[7] para transformarlo a los formatos epub y mobipocket. Esta herramienta, que es de software libre, también nos sirve para editar metadatos, añadir una portada, etc.

4. Conclusiones

En este artículo hemos presentado un proyecto para la creación de libros digitales bilingües interactivos con soporte de audio. El objetivo es facilitar la lectura en lengua extranjera y mejorar el nivel de lengua. Actualmente el mercado del libro se

[7] http://calibre-ebook.com/


[*] 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. All emotions, and that one particularly, were abhorrent to his cold, precise but admi

[*] 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. Su inteligencia fría, llena de precisión, pero admirablemente equilibrada, era en extremo opuesta a cualquier clase de emociones. Yo le considero como la máquina de razonar y de observar más perfecta que ha conocido el mundo; pero como enamorado, no habría sabido estar en su papel. Si alguna vez hablaba de los sentimientos más tiernos, lo hacía con mofa y sarcasmo. Admirables como tema para el observador, excelentes para descorrer el velo de los móviles y de los actos de las personas. Pero el hombre entrenado en el razonar que admitiese intrusiones semejantes en su temperamento delicado y finamente ajustado, daría con ello entrada a un factor perturbador, capaz de arrojar la duda sobre todos los resultados de su actividad mental. Ni el echar arenilla en un instrumento de gran sensibilidad, ni una hendidura en uno de sus cristales de gran aumento, serían más perturbadores que una emoción fuerte en un temperamento como el suyo. Pero con todo eso, no existía para él más que una sola mujer, y ésta era la que se llamó Irene Adler, de memoria sospechosa y discutible.

[*] Era poco lo que yo había sabido de Holmes en los últimos tiempos. Mi matrimonio nos había apartado al uno del otro. Mi

Figura 1: Enlace entre párrafo original y traducido

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 el potencial cliente, ni en precio, ni en funcionalidades. En nuestro proyecto tratamos únicamente con obras en dominio público, pero las editoriales que tengan los derechos de autor y de traducción de una obra se pueden plantear seriamente la posibilidad de publicar las obras en este formato. El libro digital se asemejaría a una película en DVD, donde el usuario puede 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. Actualmente disponemos de unas pocas muestras de textos paralelos en diferentes formatos que se pueden descargar de http://lpg.uoc.edu/InLector. Nuestra intención es obtener financiación externa para mejorar las herramientas de creación de los libros paralelos. También estamos abiertos a colaboraciones externas ya sea en la mejora de las herramientas como en la creación de nuevas obras.