

# InLéctor: enhanced bilingual e-books for language learning

Antoni Oliver\*, Marta Coll-Florit, Teresa Iribarren, Salvador Climent

*Universitat Oberta de Catalunya (UOC), Arts and Humanities Department*

\*Corresponding author: Tibidabo Avenue, 39-43, 08035 Barcelona, Spain, E-mail address: aoliverg@uoc.edu

Received: 2014-01-17; Accepted: 2014-05-03

## Abstract

The InLéctor project aims to promote reading in original language, offering an interactive scenario which facilitates foreign language teaching and self-learning, as well as the study of literature. In order to achieve this aim, the project develops computational techniques which provide automatic generation of bilingual e-books, incorporating dictionaries, ontologies and audios. Bilingual reading offers the option of moving from the original text to the translated one with a single click. The audio support is based on the human reading of the work in its original language. Moreover, it is planned to incorporate an interactive dictionary service which will show in advance the most difficult words in the text according to the language level of the user. Augmented information will also provide links to encyclopedic entries, images and audio directly related to the fragment of the text being read. The project ensures the creation and distribution of these parallel books in different output formats (html, epub and mobi), which are compatible with the most common tablets and e-books nowadays. The literary works and their translations are published in the public domain. The developed programs are based on free software and will also be published under a free license.

## Keywords

Language learning; e-books; translations; text alignment; text augmentation



## 1. Introduction

The main goal of the InLéctor project is to promote reading in original version, developing bilingual e-books, with text and audio. More specifically, in this project we want to encourage reading in foreign languages offering an environment of interactive reading, which will allow to know a priori the most complicated words of a chapter; links to the translated paragraphs, allowing the user to switch from the original version to the translated one immediately; and audio support corresponding to a human voice reading of the work in the original language.

This educational resource is particularly targeted at students with an intermediate / advanced foreign language level who want to read literary works in original version, and it is directly applicable in courses involving language learning or courses focusing on the study of literature.

## 2. Motivation

It is well known that the Internet and digitization has opened vast prospects for disseminating literary heritage. All over the world, a wide range of state institutions have developed large-scale digitization projects while a private company is conducting a mass digitization one (Coyle, 2006): Google Books. Thus, nowadays any citizen with any Internet device provided with a screen has the access to these entire online libraries.

In the age of globalisation, the online reading puts in contact the reader to a huge number of literary works. However, in the context of new reading experiences, which are no longer monocultural nor monolingual (Cassany, 2006), for the understanding of literary oeuvres written in different languages it remains essential the mediation of the literary translator. For that reason, InLéctor goes one step further than other digitization projects. Developed by scholars of the Universitat Oberta de Catalunya, InLéctor is an online library where the readers can find the original work, in textual and audio version, beside its translation into Spanish or Catalan. The Spanish and Catalan versions, published along the XX century, are written by literary translators.



InLéctor is a small-scale literary heritage digitization project inspired by the UNESCO Universal Declaration on Cultural Diversity (2001) and enrolled in the free culture movement (Lessig, 2004). It has been designed to favour dialogue between different literatures and facilitate intercultural understanding and exchange. It has also been conceived to enhance the interlingual process and narrow the gap between a minority language such as Catalan and hegemonic ones. On the one hand, this work in progress library aims to promote the literary canon and provide a better comprehension of the works of authors like Jane Austen, Charles Dickens, Fyodor Dostoyevsky, Alexandre Dumas or Robert Louis Stevenson for Spanish and Catalan readers. On the other hand, it approaches Spanish and Catalan texts to English, French and Russian readers.

Although all the translations are unidirectional, mainly from English, Russian or French to Spanish or Catalan, the good point of InLéctor is that with this site some Catalan historical literary translations are not only preserved but also more widely accessible. As Catalan is a source-language intensive (Cronin, 1995), this project aims to reduce the asymmetry and hierarchy in the global literary system, balancing relationships between minority and dominant languages and literatures. That is why InLéctor has been conceived as a site to promote transformation: the careful process of digitizing the bilingual texts, the tools and content provided and the chance for the users to interact will give visibility to Catalan literature in the Internet.

Moreover, it is widely accepted that reading in the original language is one of the most powerful tools for second-language acquisition (SLA). Although it seems also clear that free reading is not a sufficient source for SLA (Cobb, 2007), the input hypothesis, posed by Stephen Krashen (1985), has reached a widespread influence in language education. Such hypothesis claims that linguistic competence is mainly subconsciously acquired and it is enhanced by comprehension of language input that is slightly more advanced than the student's current level.

Consistent with this, to increase exposure of language learners to original-language written input, Krashen and many others have proposed promoting extensive reading (Day and Badford, 1998) and its practical implementation, free voluntary reading. These learning programs rely on student's motivation so that they are asked and allowed to read



books they may enjoy and to read it at their own pace without minimal or null assessment. It is claimed that such reading for pleasure eases acquisition and learning of vocabulary and idiomatic constructions. Moreover, some argue that text augmentation for difficult words is a good strategy as well for vocabulary acquisition (Rott, Williams and Cameron, 2002).

Most recently, the availability of texts in digital environments might enhance such educational potential by providing the reader with on-demand information and even interaction with the content. Automatically displayed glosses, translations or Wikipedia summaries are examples of text augmentation. The use of these instruments and other kind of digital augmentation in learning environments have given rise to a new paradigm, augmented learning, which is defined as an on-demand learning technique where the environment adapts to the needs and inputs from learners (Wang, 2012).

### **3. Functionalities**

The e-books developed within the inLéctor project will include the functionalities described below.

#### **3.1. Bilingual text**

We will develop several electronic book formats which will allow the reading of the original version of a book with direct access to its translation into another language. Specifically, the original text and the translated version will be aligned at a sentence level. Thus, if the reader wants to see the translation of a given sentence, just by clicking on the original sentence the translation will be displayed in its context, namely, within the translated work. In this way, the reader will be able to keep on reading the translation and return to the original text at any time.



### 3.2. Audio support

The electronic books will also give direct access to an audio support, which will provide human voice reading of the work in the original language. This oral dimension can be very useful to improve the comprehension and pronunciation of a foreign language. In this case, the alignment between the text and the audio will not be performed at a sentence level, but at a broader level still to be determined (paragraph or chapter).

### 3.3. Interactive glossary

Before starting the reading of a chapter or fragment of a work, the user can specify its language level, which will provide a glossary of the most difficult words in the text. The goal is to help the user to learn the meaning of these words *a priori*, in order to have a more agile and less interrupted reading. This interactive glossary will be available for download from the project website<sup>1</sup> indicating the work or chapter and the language level of the user. To generate bilingual dictionaries several free sources will be used, such as Wiktionary<sup>2</sup>, WordNets (Bond and Paik, 2012) or Apertium transfer dictionaries (Forcada, Tyers, and Ramírez, 2009).

Free bilingual and monolingual dictionaries will also be generated, so that they can be used as default dictionaries in different devices. In this way, the user will be able to look up the meaning or definition of a word at any time. The project has already generated bilingual dictionaries for the language pairs English-Spanish and English-Catalan, which are available for download from the project website.

### 3.4. Augmented Reading

The three functionalities mention so far (aligned bilingual text, human voice reading of the work and interactive glossary) are good examples of text augmentation. In addition to these functionalities, three types of augmented information will be included: 1)

---

<sup>1</sup> <http://inlector.wordpress.com/>

<sup>2</sup> <http://www.wiktionary.org/>



encyclopedic, providing information about people, places, geographical or historical events; 2) visual, offering relevant images about the text we are reading, for example art works, monuments or landscapes; and 3) sound, offering relevant musical excerpts about the text being read. The encyclopedic augmented information will come mainly from the Wikipedia<sup>3</sup>. Regarding the latter two aspects, images and sound, Wikicommons<sup>4</sup> will serve as the main source. Once all these functionalities are conjugated, a digital book becomes a multimedia product, with text, images and sound.

### **3.5. Interaction between users**

The project website will allow the interaction between users, including several interactive functionalities such as the possibility to share feedback or concerns about a particular passage of the book. This interaction will take place through the use of social networks like Facebook or Twitter.

## **4. Workflow**

### **4.1. Selection of the authors, works and languages**

The first step in the procedure of creation of an enhanced bilingual e-book is the selection of the work and the translated languages. In all cases the books are bilingual with the original language and one or more target languages. The initial working languages of this project are English, French and Russian, with translations into Spanish or Catalan. In the near future more languages will be included. In our project we are working only with works in the public domain, that is, works with no copyright, neither in the original nor in the translation, which allows us to publish the resulting e-books freely and with no restrictions. Although the expiration of copyright depends on the law of each country, in general a work can be considered in the public domain if it has been over seventy years since the death of its author.

---

<sup>3</sup> <http://www.wikipedia.org/>

<sup>4</sup> <http://commons.wikimedia.org>



The original works and their translations will be obtained primarily from Project Gutenberg<sup>5</sup> and Wikisource<sup>6</sup>. In these sources we can find electronic versions for original works in a lot of languages as well as translations into a great number of languages. Unfortunately, there is a lack of availability of translations into Catalan, one of the languages of the project. In this case, we are using OCR techniques to extract the text from paper books. In few cases, manual copying of the book is also used. The goal is to obtain a clean text file containing the original work and a file containing the translation (Figure 1).

A SCANDAL IN BOHEMIA	ESCÁNDALO EN BOHEMIA
I.	1.
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...	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...

**Figure 1.** A clean text file containing the original work and the translation

<sup>5</sup> <http://www.gutenberg.org/>

<sup>6</sup> <http://wikisource.org/>



## 4.2. Transformation to Docbook

Once we have a clean text file of the selected work we transform it into a Docbook document (Figure 2). Docbook (Walsh and Muellner, 1999) is a standard format based on XML for the representation of the logical structure of a book. The transformation of the text into a Docbook file is done because there are many freely available algorithms and modules for reading and transforming files into this format.

<chapter>	<chapter>
<title>A SCANDAL IN BOHEMIA</title>	<title>ESCÁNDALO EN BOHEMIA</title>
<section>	<section>
<title>I.</title>	<title>1.</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>	<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>

**Figure 2.** A Docbook document

The transformation from text into Docbook can be done by means of simple scripts or using the macros features of any text editor.

## 4.3. Segmentation of the text

The alignment of the books is done at sentence level. For this reason one of the first steps for the creation of the bilingual e-books is the segmentation of both original and





translated documents into sentences (Figure 3). At this moment we use a generic text segmentation from the Natural Language Toolkit (Loper and Bird, 2002), a Python library for Natural Language Processing.

A SCANDAL IN BOHEMIA	ESCÁNDALO EN BOHEMIA
<p>	<p>
I.	1.
<p>	<p>
To Sherlock Holmes she is always THE woman.	Ella es siempre, para Sherlock Holmes, la mujer.
I have seldom heard him mention her under any other name.	Rara vez le he oído hablar de ella aplicándole otro nombre.
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.
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.
...	...

**Figure 3.** Segmentation of both original and translated documents into sentences

#### 4.4. Lemmatization

In order to obtain better results in the alignment process both the original and the translated text are lemmatized, that is, all the words are reduced to their base forms (Figure 4). For the lemmatization step we are using Freeling (Padró and Stanilovsky 2012) and TreeTagger (Schmid 1994), depending on the languages. The goal is to reduce the number of word forms to make the task easier for the statistical sentence alignment algorithm.



A SCANDAL IN BOHEMIA	ESCÁNDALO EN BOHEMIA
<p>	<p>
I.	1.
<p>	<p>
to Sherlock Holmes she be always the woman .	ella ser siempre para Sherlock Holmes el mujer .
I have seldom hear him mention her under any other name .	raro vez él haber oír hablar de él aplicar otro nombre .
in his eye she eclipse and predominate the whole of her sex .	a el ojo de Sherlock Holmes eclipsa y sobrepasar a todo suyo sexo .
it be not in he feel any emotion akin to love for Irene Adler .	no ser que haber sentido por Irene Adler nada que se parecer al amor .

**Figure 4.** Lemmatization

#### 4.5. Alignment

For the automatic alignment of the texts at sentence level we are using Hunalign (Varga et al., 2005). This program is able to achieve very good results in our task and in most of the cases all the proposed alignments are correct. The program allows very quick alignments as it is able to align a full book in seconds. The alignment is done using the lemmatized version of the books (Figure 5).

1.56762 A SCANDAL IN BOHEMIA	ESCÁNDALO EN BOHEMIA
1.89878 I 1	
1.45135 to Sherlock Holmes she be always the woman .	ella ser siempre para Sherlock Holmes el mujer .



1.21031 I have seldom hear him mention her under any other name . raro vez él haber oír hablar de él aplicar otro nombre .

0.906857 in his eye she eclipse and predominate the whole of her sex . a el ojo de Sherlock Holmes eclipsa y sobrepasar a todo suyo sexo .

1.008 it be not in he feel any emotion akin to love for Irene Adler . no ser que haber sentido por Irene Adler nada que se parecer al amor .

**Figure 5.** Text alignment using the lemmatized version of the books

However, having this alignment it is straightforward to obtain the alignment of the files with full word forms (Figure 6).

1.56762 A SCANDAL IN BOHEMIA ESCÁNDALO EN BOHEMIA

1.89878 I I

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.

**Figure 6.** Alignment of the files with full word forms



To improve even more the results of the alignment process the program can be provided with a bilingual dictionary.

#### 4.6. Creation of the bilingual html

The creation of the bilingual html file is done by simple scripts that use the Docbook versions of the original, the translation and the alignment file (Figure 7). The script adds links between the original and the translated sentences. When some original sentence does not have a translated sentence in the alignment file, the link leads to the nearest translated sentence.

```
<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>
```

Figure 7. Bilingual html file

#### 4.7. Creation of the bilingual e-books

For the transformation of the html into epub and mobi files we are using the scripts provided with the Calibre tool. After the transformation some manual editions should be made in order to change the style files to avoid the underline of the sentences holding a link, as most of the sentences in the e-book contain those links.



#### **4.8. Translation memories**

As a secondary product of the InLéctor project we are collecting all the alignments and creating translation memories. When the number of published bilingual e-books will be big enough, these translation memories will be useful for literary translation studies, as the different solutions to several translation problems will be easily found in the translation memories. These translation memories will be freely distributed and will be also published in a freely searchable database.

#### **4.9. Audio support**

In the following months we will start to publish bilingual e-books that include the audio corresponding to the human reading in the original language. To do so we will use the new features of epub 3 format that allows the insertion and synchronization of audio using the <audio> tag of HTML 5. The main source of the audio corresponding to the reading of the original works will be Librivox<sup>7</sup>, a project in which a large number of volunteers read chapters of books that are copyright free and publish audio files also in public domain.

### **5. Conclusions**

Multiple and too often apocalyptic debates about the battle between paper and e-books, the Internet piracy or the decay of the culture of literary reading among university students are still in the center of the public discussion. In front of these sterile debates, the scholars involved in the InLéctor project are connecting technology and humanistic knowledge in order to provide to active readers an attractive and comprehensive way to both learn languages and enjoy classic literature.

Nowadays, learners use mobile devices such as smartphones, e-readers and tablets to read, write and listen. They are used to access to the content with a single click or

---

<sup>7</sup>

<http://librivox.org/>



touchscreen and they like to share any kind of experiences, like the reading one, in social networks like Facebook or Twitter. The InLéctor site has been specially designed for this new type of learners. In this small-scale literary heritage digitization project they will find selected oeuvres in the original version and their translations to Spanish or Catalan within an augmented reality scenario. With a single click or touchscreen, they will be able to choose the way they want to enjoy these oeuvres: reading the original text, reading the translated version or listening an actor or a rhapsode reading or performing the literary work.

Moreover, InLéctor will be a virtual meeting point where citizens from different literary systems will have the opportunity to share reading experiences through social networks while they learn languages. It will be also a useful tool for literary translators, because the site will provide translation memories, and it can even become a forum of textual discussion and exchange. Finally, as InLéctor has the aim to preserve and make more widely accessible Catalan historical literary translations, the project contributes to approach and balance minority and dominant literatures.

## 6. References

- Bond, F. and Paik, K. (2012). A survey of wordnets and their licenses. In Proceedings of the 6th Global WordNet Conference, GWC 2012 (pp. 64–71).
- Cassany, D. (2006). *Rere les línies. Sobre la lectura contemporània*. Barcelona: Empúries.
- Cobb, T. (2007), Computing the Vocabulary Demands of L2 Reading, *Language Learning & Technology* 11 (3), 38–63.
- Coyle, K. (2006). Mass digitization of books, *Journal of Academic Librarianship*, 32 (6). <<http://www.kcoyle.net/jal-32-6.html>> (Accessed 07.02.2014)
- Cronin, M. (1995). *Altered States: Translation and Minority Languages*, *TTR: traduction, terminologie, rédaction*, 8 (1), 85-103.
- Day, R. and Bamford, J. (1998). *Extensive Reading in the Second Language Classroom*. Cambridge: Cambridge University Press.



Forcada, M. L., Tyers, F. M. and Ramírez, G. (2009). The Apertium machine translation platform: five years on. In Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation (pp. 3–10).

Krashen, S. D. (1985). The input hypothesis: Issues and implications. London, NYC: Longman.

Lessig, L. (2004). Free Culture. New York: The Penguin Press.

Loper, E and Bird, S. (2002). NLTK: The Natural Language Toolkit. In Proceedings of the ACL-02 Workshop on effective tools and methodologies for teaching natural language processing and computational linguistics. Volume 1. Association for Computational Linguistics. Stroudsburg, PA, USA (pp. 63-70).

Padró, L and Stanilovsky, E. (2012) FreeLing 3.0: Towards Wider Multilinguality. In Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul, Turkey.

Rott, S., Williams, J. and Cameron, R. (2002). The effect of multiple-choice L1 glosses and input-output cycles on lexical acquisition and retention, *Language Teaching Research* 6(3), 183–222.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L. and Trón, V. (2005). Parallel corpora for medium density languages. In Proceedings of the RANLP 2005 (pp. 590–596).

Walsh, N. and Muellner, L. (1999). DocBook: The definitive guide. O'Reilly Media

Wang, X. (2012). Augmented Reality: A new way of augmented learning. *eLearn Magazine*. <https://elearnmag.acm.org/archive.cfm?aid=2380717> (Accessed 04.02.2014)

