InLéctor: automatic creation of bilingual e-books

Antoni Oliver González

revista.tradumàtica
tecnologies de la traducció

Antoni Oliver González
Universitat Oberta de Catalunya
aoliverg@uoc.edu;
ORCID:
0000-0001-8399-3770

Abstract

In this paper, a system for the automatic creation of parallel bilingual electronic books is presented. The system allows creating e-books, where source sentences are linked with the corresponding target sentences. Users can read in the original, and clicking on a given sentence, the corresponding sentence in the target language is shown. Then she or he can continue reading the translation and coming back to the original version clicking in a target language sentence. The source language book is automatically aligned at the sentence level with the target language translation of the book. This system is not using a machine translation system, but instead, it shows the published translation of the original work in the given target language. We have created several bilingual e-books using classic novels and its translations in the public domain, but the same system can be used for any book, provided you have the rights for the original and the translation. The system is aimed to people willing to read in the original, having a mid-high level of the source language. We also present the process of creation of bilingual dictionaries from free lexical resources. Both resources, the bilingual e-book and the bilingual dictionary can be of great help for readers willing to read books in the original version.

Keywords: e-books; parallel texts; reading aid.

Resum

Aquest article presenta un sistema de creació automàtica de llibres bilingües amb textos alineats. El sistema permet crear llibres electrònics en què l'oració en la llengua de partida està vinculada amb la corresponent oració en la llengua d'arribada. Els usuaris poden llegir en la llengua original i veure l'oració corresponent en la llengua d'arribada fent clic sobre l'oració de partida. D'aquesta manera pot continuar llegint en la llengua d'arribada o tornar a la versió original fent clic sobre qualsevol oració. El llibre en la llengua de partida s'alinea automàticament a nivell d'oració amb la seva traducció en la llengua d'arribada. Aquest sistema no utilitza la traducció automàtica sinó que, per contra, mostra la traducció publicada de l'obra original en la llengua d'arribada. Hem creat diversos llibres electrònics bilingües utilitzant novel·les clàssiques i la seva traducció en domini públic, però aquest sistema també és utilitzable per a qualsevol llibre per al qual l'usuari tingui drets sobre l'original i la traducció. El sistema s'adreça a persones que vulguin llegir

Rebuda: 12 de juny de 2017 / Acceptació: 11 de novembre de 2017 / Publicació avançada: 23 de desembre de 2017

originals en una llengua en què tenen un nivell mitjà o alt. Aquest article també presenta el procés de creació de diccionaris bilingües a partir de recursos lèxics gratuïts. Tots dos recursos, el del llibre electrònic i el de diccionaris bilingües, poden ser de gran ajuda per a tots aquells lectors que vulguin accedir a la versió original dels llibres.

Paraules clau: llibre electrònic; textos paral·les; suport a la lectura.

Resumen

Este artículo presenta un sistema de creación automática de libros bilingües con textos alineados. El sistema permite crear libros electrónicos en los que la oración en la lengua de partida está vinculada con la correspondiente oración en la lengua de llegada. Los usuarios pueden leer en la lengua original y ver la oración correspondiente en la lengua de llegada haciendo clic sobre la oración de partida. De esta manera puede continuar leyendo en la lengua de llegada o volver a la versión original haciendo clic sobre cualquier oración. El libro en la lengua de partida se alinea automáticamente a nivel de oración con su traducción en la lengua de llegada. Este sistema no utiliza la traducción automática, sino que, por el contrario, muestra la traducción publicada de la obra original en la lengua de llegada. Hemos creado diversos libros electrónicos bilingües utilizando novelas clásicas y su traducción en dominio público. Sin embargo, este sistema también se puede utilizar para alinear cualquier libro del que el usuario disponga de derechos sobre el original y la traducción. El sistema va dirigido a persona que quieran leer originales en una lengua en la que tienen un nivel medio o alto. Este artículo también presenta el proceso de creación de diccionarios bilingües a partir de recursos léxicos gratuitos. Ambos recursos, el del libro electrónico y el de diccionarios bilingües, pueden ser de gran ayuda para todos aquellos lectores que deseen acceder a la versión original de los libros.

Palabras clave: libro electrónico; textos paralelos; apoyo a la lectura.

1. Introduction

Many people have studied some foreign languages, achieving an intermediate or high level of proficiency, but they don't use this foreign language in her or his daily life. Some of these people like reading, and they would like to read books in the original. When reading, they realize that they need to search unknown words in dictionaries, and even doing so, sometimes it's hard to understand the whole sentence, so being difficult to follow the book. After some reading sessions, most of these readers withdraw the book and start to read a book in her or his mother tongue. Some studies (Nuttall, 1996) highlight the importance of reading in the study of a foreign language. There has been a long tradition in the publication of *parallel books*, where the original was in one side, and the translation on the other side. With these parallel books, the reader can choose to read the original or the translation, and to switch from one to the other simply looking at the left or at the right. This kind of books can also receive the name of *bilingual books* (Ernst-Slavit and Mulhern, 2003). Since bilingual books involve translations from one language to another, (Semingson et al., 2015) point out that the quality of translation is an important consideration. For this reason in such bilingual books we use a human translation, usually also published as an independent book, and we avoid the use of machine translation systems. For the InLéctor collection we are using original works and translations in the public domain,

avoiding intellectual property restrictions. The same processes can be applied to any books and translations by their copyright holders.

Extensive reading, also known as *for pleasure reading* (Bamford and Day, 1997), in a tension-free environment can be an effective activity for the enhancement of learner's language skills, both receptive and productive (Hafiz and Tudor, 1989). This same idea is found in (Prowse, 2002) which states, citing (Krashen, 2004), that *free voluntary reading* is an important activity to get the benefits from extensive reading. Some researchers (Chin-Neng et al., 2013) have found that the integration of e-books in extensive reading activities helps to improve students' reading attitude, reading comprehension and vocabulary. The books in the InLéctor collection are intended for the promotion of reading in the original language. They can be used as learning material in language courses. They can also be freely used by readers willing to practice and improve their foreign language skills.

New reading devices, as e-books and tablets, are the perfect platform for bilingual reading. On one hand, they allow automatic query to dictionaries (both installed in the same device or accessible through Internet). Using this functionality, readers can consult the meaning of a word or expression with a couple of clicks. On the other hand, the ability of the different formats for e-book publishing to incorporate links, allows relating one paragraph or sentence with its translation. When parallel books are published in paper, they need the double of space and weight, being that a big drawback for large novels to be published in a bilingual edition. With electronic formats this problem is solved, as the only limit is the memory of the device.

In this paper, we present the process of automatic creation of bilingual e-books using free software tools. We also present the collection InLéctor [1], where our bilingual books are published. There are other initiatives, both free and commercial, publishing bilingual e-books, as for example:

- Doppeltext[2]: it's a commercial company that provides the books in several formats. For example, for iBooks in iPad source and target texts are presented face-to-face; for electronic books clicking in a sentence takes you to the translated version and when reading in a web browser, clicking in a sentence opens a pop-out with the translation.
- FarkasTranslations[3]. Provides several free bilingual e-books and a program, Ebookmaker, available for Windows only, to create bilingual e-books. This project also accepts suggestions for new publications and allows volunteers to join and create new books.
- Bilinguis[4]: provides free bilingual e-books than can be read online. Original and translation are presented face-to-face.

[1] InLector, https://inlector.wordpress.com/
[2] Doppeltext, http://www.doppeltext.com/
[3] FarkasTranslations, http://www.farkastranslations.com/bilingual_books.php
[4] Bilinguis, http://bilinguis.com/

The InLéctor project has been developed under a grant for Teaching Innovation Projects Aplica 2013 from the Universitat Oberta de Catalunya (UOC).

This paper is organised as follows. In section 2 we present several sources for books in the public domain. In section 3 all the steps for the creation of bilingual e-books are explained in detail. We also present the tools and instructions for the creation of bilingual e-books. As all the tools hold a free software licence, any person can create its own bilingual e-books following these instructions. In section 4 the tools and instructions for the creation of bilingual dictionaries from several free lexical resources are presented. In section 5 a derived product for this project, namely the multilingual translation memories, are presented. We also discuss the utility of using these translation memories in literary translation. In section 6 we present the bilingual e-books published so far in the InLéctor collection. And lastly, the conclusions and future work are presented in section 8.

2. Collections of books in the public domain

2.1 The Project Gutenberg

The aim of Project Gutenberg is to digitize and archive cultural works, mainly full text of books in the public domain. Now the project provides around 53,000 free e-books in several formats, including plain text, *html*, *epub* and *mobi*. The project is supported by thousands of volunteers that proofread the digitized books. In table 1 we can observe the number of books for the top 10 languages. Unfortunately, the catalogue only gives the author, title and language and there is not an easy way to detect books in the original language along with the available translation. For the moment, this task must be done manually each time we are interested in a given book.

Language	Books
English	30,768
French	1,834
German	776
Finnish	592
Dutch	536
Portuguese	516
Chinese	405

Spanish	300
Italian	291
Greek	147

Table 1: Number of books for the 10 top languages in Project Gutenberg (source: Project Gutenberg)

Language	Articles
French	2,077,150
English	1,744,467
Bengali	547,808
Russian	451,896
Polish	447,559
Tamil	409,920
German	389,269
Italian	277,046
Hebrew	192,456
Spanish	169,855

Table 2: Number of articles for the 10 top languages in Wikisource (source: Wikisource)

2.2 Wikisource

Wikisource[5] is a Wikimedia Foundation project that aims to create a library of texts in source language and translations to other languages. The documents published in Wikisource must hold a free licence or be in the public domain. The works are published in the mediawiki format, but they can be downloaded in other formats as *pdf*, *epub* and *mobi*. Wikisource is a multilingual project, and source texts and translations are linked, allowing easy access to the same text in several languages. In table 2 figures of the size of Wikisource for the 10 top languages are presented. Please, note that figures indicate the number of articles, not the number of books. An article can be a chapter of a book, as well as general information about the author or a given book. For example, for Anatole France's "L'Île des Pingouins" we can find:

- An article for the author[6]

[5]Wikisource, https://wikisource.org

[6]Anatole France, https://fr.wikisource.org/wiki/Auteur:Anatole_France

- An article for the catalogue of books of the author[7]
- An article for the book[8] with the table of contents
- As the book is divided in 8 parts, we have an article for each part, containing all the chapters of the part[9].

This structure can be different for other books. If the translated version to a given language is available, in each of these articles we would have a link to the translated content. Given this variability in the structure of a book inside Wikisource, if we want to get the full text of the book from this source, we would need to manually copy the text of each page into a text file for further processing.

3. Creation of bilingual e-books

In this section, the process of creation of bilingual e-books is presented. All the examples are taken from the one book of the collection, the French-Catalan version of the Anatole France's novel *L'Île des Pingouins*. The starting point is getting the text versions of the novel in the source and target languages. We usually use collections of books in the public domain as explained in the previous section. We have obtained the French version from Project Gutenberg; but the Catalan translation is an unpublished one, done by J.F. Vidal Jové, and given to our project by his heirs. A full edition process has been performed to this translation (Iribarren et al., 2016), and we use this edited text to create the bilingual edition.

To facilitate the process of creation of the bilingual e-books, a full set of scripts written in Python has been developed. For the moment, no graphical user interface is available and all scripts have to be run from the terminal. With the help of these scripts, we can speed up the process. Once obtained the source and target texts, the full process of creation of the bilingual books takes less than a couple of hours, on average. The creation process is divided in five steps:

1/ Conversion to DocBook
2/ Segmenting and numbering the segments in the DocBook files
3/ Text alignment
4/ Creation of the bilingual DocBook
5/ Conversion to popular formats: *epub*, *mobi* and *html*

More details on the different steps are given in the following subsections.

3.1 Conversion to DocBook

Using one of the scripts and a text processor, we convert the text files to DocBook, an XML-based format for the representation of documents. As it is a standard format,

[7] Anatole France, Catalogue des œuvres, https://fr.wikisource.org/wiki/Anatole_France/Catalogue_des_œuvres
[8] "L'^Ile_des_Pingouins", https://fr.wikisource.org/wiki/L'^Ile_des_Pingouins
[9] "L'^Ile_des_Pingouins", Livre_premier, https://fr.wikisource.org/wiki/L'^Ile_des_Pingouins/Livre_premier

a great number of utilities for the processing and conversion from this format to several other formats are available. The conversion to DocBook requires a manual revision of the process in order to mark some elements, as the Python script the tag <para> at the beginning and end of all paragraphs. The script is *txt2docbook.py* and if we run it with the parameter -h the help is shown:

```
usage: txt2docbook.py [-h] -i INPUT_FILE -o OUTPUT_FILE [-l LANGUAGE]
A script to convert a txt file into a docbook file. First line must be the
title of the book, second line the name of the author and third line the
suname of the author. Parts titles: = Chapter titles: ==
optional arguments:
  -h, --help            show this help message and exit
  -i INPUT_FILE, --input_file INPUT_FILE
                        The txt input file to convert
  -o OUTPUT_FILE, --output_file OUTPUT_FILE
                        The dockbook output file
  -l LANGUAGE, --language LANGUAGE
                          The language of the book. Default eng
```

The script expects that the first line of the text file contains the title of the book, the second line the author's name and the third line the author's surname. In order to mark the parts and chapters of the book, part titles must start and end with "=", and chapter titles must start and end with "==". For example, the French text file should start as follows:

```
L'Île Des Pingouins
Anatole
France
=PRÉFACE=
Malgré la diversité apparente des amusements qui semblent m'attirer, ma
vie n'a qu'un objet. Elle est tendue tout entière vers l'accomplissement
d'un grand dessein. J'écris l'histoire des Pingouins. J'y travaille
assidument, sans me laisser rebuter par des difficultés fréquentes et qui,
parfois, semblent insurmontables.
   ...
```

And the Catalan text should start as follows:

```
L'ILLA DELS PINGÜINS
ANATOLE
FRANCE
=PREFACI=
   Malgrat la diversitat aparent dels entreteniments que sembla que
m'atreuen, la meva vida no té més que un objecte. Tota ella està bolcada
vers l'acompliment d'una gran dèria. Escric la història dels pingüins. Hi
treballo assíduament, sense deixar-me descoratjar per les sovintejants
dificultats que, a vegades, semblen insuperables.
```

If we have the French text in the file *lip-fra.txt* and the Catalan text in the file *lip-cat.txt* we can convert them into DocBook running this script two times.

```
python3 txt2docbook.py -i lip-fra.txt -o lip-fra.docbook -l fra
python3 txt2docbook.py -i lip-cat.txt -o lip-cat.docbook -l cat
```

Here we can observe a fragment of the original text:

```
<title>VIDA DE SANT MAËL</title>
<para>
<phrase id='225'>Maël, eixit d'una família reial de Cambrie, entrà quan
tenia nou anys a l'abadia d'Yvern per estudiar-hi lletres sagrades i
profanes.</phrase>
<phrase id='226'>A l'edat de catorze anys renuncià a la seva herència i féu
vots de servir el Senyor.</phrase>
<phrase id='227'>Repartia les seves hores com comana la Regla, entre el
cant dels himnes, l'estudi de la gramàtica i la meditació de les veritats
eternes.</phrase>
</para>
```

and the same process is performed in the translated text:

```
<title>VIDA DE SANT MAËL</title>
<para>Maël, eixit d'una família reial de Cambrie, entrà quan tenia nou anys
a l'abadia d'Yvern per estudiar-hi lletres sagrades i profanes. A l'edat de
catorze anys renuncià a la seva herència i féu vots de servir el Senyor...
</para>
```

3.2 Segmenting and numbering the segments in the DocBook files

The next step involves the segmentation of the paragraphs into sentences. The process of segmentation consists in splitting the text into segments or similar units. This process is not trivial, as in most languages sentences are divided by dots (.), but this same character is also used in other units (as abbreviations and acronyms, for example). We can consider the following example:

...le comte Cléna, M. de la Trumelle, M. Bigourd et quelques riches dames israélites.

That should be kept together in one segment. If we use the dot as an end symbol for a segment, it would lead to the incorrect segmentation:

```
...le comte Cléna, M.
de la Trumelle, M.
Bigourd et quelques riches dames israélites.
```

To perform the segmentation process, we use the Natural Language Toolkit (NLTK)[10] (Bird, 2006), a Python library providing functionalities for several Natural Language Processing (NLP) related tasks. NLTK provides pre-trained segmenters for 17 languages: Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Italian, Norwegian, Polish, Portuguese, Slovene, Spanish, Swedish and Turkish. It also provides tools for training segmenters for other languages, based on (Kiss and Strunk, 2006). This algorithm is based on the idea that most ambiguities in sentence boundaries can be avoided once abbreviations have been identified, so the algorithm learns abbreviations from raw text, taking into account three basic characteristics of abbreviations:

- Abbreviations can be seen as collocation consisting of a truncated word and a final period.
- Abbreviations are usually short.
- Abbreviations sometimes contain internal periods.

Authors report slightly worse precision results than other systems, but this algorithm has the great advantage that can be easily adapted to a new language or domain, as it only needs raw text to be trained. We have used these tools for training a segmenter for Catalan using the script *trainsegmenter.py*. If we run this script with the "-h" parameter, the help is shown:

```
python3 trainsegmenter.py -h
usage: trainsegmenter.py [-h] -i INPUT_FILE [-a ABBREVIATIONS] -s
SEGMENTER
A script to train a segmenter using a book in docbook file. A list of
abbreviations should be also given.
optional arguments:
  -h, --help            show this help message and exit
  -i INPUT_FILE, --input_file INPUT_FILE
                        The docbook input file to learn from.
  -a ABBREVIATIONS --abbreviations ABBREVIATIONS
                        The list of known abbreviations.
  -s SEGMENTER, --segmenter SEGMENTER
                           The learnt segmenter.
```

Using the *abr-fra.txt* and *abr-cat.txt* files containing some French and Catalan abbreviations we can train the segmenters using the script:

[10]Natural Language Toolkit, http://nltk.org

```
   python3 trainsegmenter.py -i lip-fra.docbook -a abr-fra.txt -s
segmenter-fra.pickle
   python3 trainsegmenter.py -i lip-cat.docbook -a abr-cat.txt -s
segmenter-cat.pickle
```

The French segmenter is stored in the file *segmenter-fra* and the Catalan segmenter is stored in the file *segmenter-cat*

Now, to segment the DocBook files we can use the script *segmentdocbook.py*. If we run this script with the "-h" parameter, the help is shown:

```
python3 segmentdocbook.py -h
usage: segmentdocbook.py [-h] -i INPUT_FILE -s SEGMENTER -o OUTPUT_FILE -a
                         HUNALIGN_FILE
A script to segment docbooks.

optional arguments:
  -h, --help            show this help message and exit
  -i INPUT_FILE, --input_file INPUT_FILE
                        The docbook input file to learn from.
  -s SEGMENTER, --segmenter SEGMENTER
                        The learnt segmenter.
  -o OUTPUT_FILE, --output_file OUTPUT_FILE
                        The numbered docbook output file.
  -a HUNALIGN_FILE, --hunalign_file HUNALIGN_FILE
                        The output file for Hunalign.
```

To create the segmented and numbered DocBook s for the French and Catalan texts we run the script as follows:

```
python3 segmentdocbook.py   -i lip-fra.docbook -o lip-fra-num.docbook -s
segmenter-fra.pickle -a lip-fra.hunalign
python3 segmentdocbook.py   -i lip-cat.docbook -o lip-cat-num.docbook -s
segmenter-cat.pickle -a lip-cat.hunalign
```

After the segmentation process, source text segments are numbered and the prefix *ss* (source segment) is added. The prefix *ts* (target segment) is added to the number of the target text segments. Here, we can observe the segmented and numbered DocBook corresponding to the original text.

```
<title>VIE DE SAINT MAËL</title>
<para>
<phrase id='236'>Maël, issu d'une famille royale de Cambrie, fut envoyé dès
sa neuvième année dans l'abbaye d'Yvern, pour y étudier les lettres sacrées
et profanes.</phrase>
<phrase id='237'>À l'âge de quatorze ans, il renonça à son héritage et fit
voeu de servir le Seigneur.</phrase>
<phrase id='238'>Il partageait ses heures, selon la règle, entre le chant
des hymnes, l'étude de la grammaire et la méditation des vérités éternelles.
</phrase>
</para>
```

And here the one corresponding to the target text:

```
<title>VIDA DE SANT MAËL</title>
<para>
<phrase id='225'>Maël, eixit d'una família reial de Cambrie, entrà quan
tenia nou anys a l'abadia d'Yvern per estudiar-hi lletres sagrades i
profanes.</phrase>
<phrase id='226'>A l'edat de catorze anys renuncià a la seva herència i féu
vots de servir el Senyor.</phrase>
<phrase id='227'>Repartia les seves hores com comana la Regla, entre el
cant dels himnes, l'estudi de la gramàtica i la meditació de les veritats
eternes.</phrase>
</para>
```

Note that the numbering of the source segment and the target segment does not match. In fact, the source segment *236* corresponds to the target segment *225*. The reason can be multiple: some source segments translated as a several target segments; some parts of the source text not translated in the target text; some explanatory notes only present in one of the versions, etc.

Also, note that titles don't hold a numbered phrase tag. This is to avoid linking source titles with target titles, as when creating the e-books, automatic links to table of content will be created. Linking source and target titles would lead to links with double end, causing errors in the e-book readers.

When running the scripts we are also using the parameter "-a" followed by a file name. This will create the required files for the alignment with Hunalign as it is explained in the following subsection.

3.3 Text alignment

For the automatic alignment of the original and translated text we use Hunalign (Varga et al., 2007). This program performs a fully automatic alignment of the source and target text at the sentence level. Hunalign uses a dictionary-based translation model exploiting a bilingual dictionary. Thus, it's compulsory to provide a bilingual dictionary, and if it is not available, an empty file can be used. If the bilingual dictionary is empty or it does not contain enough information Hunalign tries to automatically construct a statistical bilingual dictionary. To do so, Hunalign builds alignments using a simple similarity measure based on sentence length and the ratio of identical tokens. This is

especially interesting in texts containing a high percent of numeric expressions, as the similarity on numerical tokens is taken into account.

If no bilingual dictionary is available, as the dictionary file is compulsory, an empty file should be used. From the InLéctor project page bilingual dictionaries for several language pairs can be downloaded. These dictionaries have been built using the transfer dictionaries of the Apertium machine translation system[11] (Forcada et al., 2011). These dictionary files have the following format (we show the French - Catalan alignment dictionary):

```
Abad @ Abad
humiliant @ abaissant
rebaixat @ abaissé
descens @ abaissement
baixar @ abaisser
abandó @ abandon
abandonar @ abandoner
```

Note that, by convention, the target language comes first in these dictionaries. In table 3 we can observe the size in number of entries of the alignment dictionaries.

Language pair	Entries
Catalan - English	33,410
Catalan – Spanish	31,140
Catalan – French	12,298
Catalan – Italian	8,068
English – Spanish	28,334
French – Spanish	22,201
Spanish – Galician	25,569
Spanish – Italian	10,456
Spanish – Romanian	17,552
Spanish – Portuguese	9,940

Table 3: Number of entries of the alignment dictionaries for Hunalign (source: author)

In order to use Hunalign, we need to export the segmented and numbered DocBook files into the text file format required by Hunalign. This can be achieved in the same step when the segmented and numbered DocBook files are created, using the *segmentdocbook.py* script, as explained in the previous section. These files consist in a segment per line, and a `<p>` mark for paragraphs. Here we can observe the text of our example in this format:

[11] Apertium: http://www.apertium.org

```
<p>
Maël, issu d'une famille royale de Cambrie, fut envoyé dès sa neuvième
année dans l'abbaye d'Yvern, pour y étudier les lettres sacrées et profanes
À l'âge de quatorze ans, il renonça à son héritage et fit voeu de servir le
Seigneur.
...
<p>
```

And the Catalan translation:

```
<p>
Maël, eixit d'una família reial de Cambrie, entrà quan tenia nou anys a
l'abadia d'Yvern per estudiar-hi lletres sagrades i profanes.
A l'edat de catorze anys renuncià a la seva herència i féu vots de servir
el Senyor.
...
<p>
```

Please, note again that text marked as titles are not present in the files to align, as these elements will not carry alignment information to avoid the double link problem, as mentioned above.

To align the source (*lip-fra.hunalign*) and the target (*lip-cat.hunalign*) books using the French-Catalan alignment dictionary (*hunalign-fr-ca.dic*), we run the following command:

```
hunalign -realign -utf hunalign-fr-ca.dic lip-fra.hunalign lip-
cat.hunalign > ali.txt
```

Hunalign returns a file (*ali.txt*) with the alignments giving the source segment number, the target segment number and a score, This score indicates the confidence of the alignment. The higher this score, the better the alignment is. For our text examples, the alignment file contains:

```
236    225    1.79565
237    226    1.33817
```

This indicates, for example, that source segment number 236 corresponds to the target segment 225, with a confidence score of 1.79595.

Hunalign achieves very good values of precision. If errors are encountered, they are usually due to segmentation problems. To further improve the performance of the alignment step, a lemmatized version of the books can be used. A script that call the Freeling analyzer (Padró and Stanilovsky, 2012) converts the segmented books in Hunalign format to its lemmatized version. Lemmatizing has the effect of reducing the number of different word forms, making the alignment process easier. Also, as the bilingual dictionaries contain lemmata and not all the wordforms, the use of the

dictionaries in the alignment process is more efficient. As the number of the segments in the full form and lemmatized versions of the book are the same, the alignment file obtained for the lemmatized books can be also used for the full form version of the books. Here we can observe the text of the example in lemmatized form:

The French version:

```
<p>
Maël , issir de un familie royal de Cambrie , être envoyer des son 9 année
dans l' abbaye d' Yvern , pour y étudier les lettre sacrer et profane .
À l' âge de quatorze an , il renoncer à son héritage et faire voeu de
servir le Seigneur .
...
<p>
```

And the Catalan translation:

```
<p>
Maël , eixir d' una família reial de Cambrie , entrar quan tenir nou any a
l' abadia d' Yvern per estudiar hi lletra sagrada i profana .
a l' edat de catorze any renunciar a el seu herència i fer vot de servir el
Senyor .
…
<p>
```

As we have already said, this lemmatization step is not usually performed, as the regular process usually achieves very good results.

3.4. Creation of the bilingual DocBook

To create the bilingual DocBook the alignment file presented in the previous section will be used. In the bilingual DocBook, both source and target numbered DocBooks are merged and links between source segments and target segments are added. To create the bilingual DockBook the *create_bilingual_ebook.py* script can be used. If we run this script with the "-h" parameter the help is shown:

```
python3 create_bilingual_docbook.py -h
usage: create_bilingual_docbook.py [-h] -s SOURCE -t TARGET [-a ALIGNMENT]
                                   [-o OUTPUT]
A script for the creation of bilingual docbooks.
optional arguments:
  -h, --help            show this help message and exit
  -s SOURCE, --source SOURCE
                        The source language numbered docbook.
  -t TARGET, --target TARGET
                        The target language numbered docbook
  -a ALIGNMENT, --alignment ALIGNMENT
                        The alignment ladder file.
  -o OUTPUT, --output OUTPUT
                           The output file.
```

In our example, for the creation of the bilingual DocBook, the script should be used as follows:

```
python3 create_bilingual_docbook.py -s lip-fra-num.docbook -t lip-cat-num.docbook -a ali.txt -o lip-fra-cat.docbook
```

Here we can see the fragments corresponding to the example:

```
<title>VIE DE SAINT MAËL</title><title>VIE DE SAINT MAËL</title>
<para>
<phrase id="ss-236"><link linkend="ts-225">Maël, issu d'une famille royale
de Cambrie, fut envoyé dès sa neuvième année dans l'abbaye d'Yvern, pour y
étudier les lettres sacrées et profanes.</link></phrase>
<phrase id="ss-237"><link linkend="ts-226">À l'âge de quatorze ans, il
renonça à son héritage et fit voeu de servir le Seigneur.</link></phrase>

….
<title>VIDA DE SANT MAËL</title>
<para>
<phrase id="ts-225"><link linkend="ss-236">Maël, eixit d'una família reial
de Cambrie, entrà quan tenia nou anys a l'abadia d'Yvern per estudiar-hi
lletres sagrades i profanes.</link></phrase>
<phrase id="ts-226"><link linkend="ss-237">A l'edat de catorze anys
renuncià a la seva herència i féu vots de servir el Senyor.</link></phrase>
…
```

Please, note that as already mentioned, titles do not have links between source and translation. In the example we can observe how the source segment 236 is linked to the target segment 225; and how the target segment 225 is linked to the source segment 236.

3.5. Conversion to popular formats: epub, mobi and html

3.5.1. Conversion to epub

The *epub*[12] format is a technical standard published by the International Digital Publishing Forum (IDPF). An *epub* file is in fact a *zip* archive that contains *html* files, images, *css* style sheets, and other files, as well as metadata. A typical *epub* 2 file would have the following structure of directories (in upper case letters) and files (in lower case letters):

- META-INF
- content.xml
- mimetype
- OEBPS

[12]EPUB, http://idpf.org/epub

- content.opf
- stylesheet.css
- toc.ncx
- xaa.html
- xab.html
- ...

Epub 3 is the latest version of this format and is based on HTML5 standard. This means that *epub* 3 publications can contain audio, video and interactivity. This new version is still not fully implemented in all reading devices. When this version will be widely implemented in popular reading devices, we will include new functionalities in the InLéctor collection. The most relevant one will be the inclusion of synchronized audio version of the books. The user will be able to activate the audio corresponding to the human reading of the book, and the sentences in the book will be highlighted while the sound corresponds to the text. The audio of the books will be taken from Librivox[13], a project providing free public domain audiobooks in several languages, read by volunteers.

To convert the bilingual DocBook to *epub* we use the program *dbtoepub*[14]. This tool will do all the conversion, but once done, if we open the resulting *epub* we will notice that all the sentences are underlined, as it is the default option in *epub*. To avoid the underlined links me must include a stylesheet.css in the OEBPS directory of the *epub* file specifying that the links shouldn't be underlined:

```
.link {\\
    color: \#000;
    cursor: inherit;
    line-height: 1.2;
    text-decoration: none
    }\\
```

This operation can be done in different ways, for example, using Calibre[15] to edit the *epub* and add the stylesheet. Calibre has a convenient feature that adds the reference to the stylesheet in all the required html files. To avoid the use of an external editor, we can tell to *dbtoepub* to use this specific stylesheet.css:

```
dbtoepub -c stylesheet.css lip-fra-cat.docbook
```

This will create the *epub* file *lip-fra-cat.docbook.epub* that can be read in any electronic book device.

[13]Librivox, https://librivox.org/
[14]*dbtoepub* is Available at http://docbook.sourceforge.net
[15]Calibre, http://calibre-ebook.com/

3.5.2. Conversion to mobi

The *mobi* format is mainly used by the Amazon's Kindle e-book readers. As these readers are very popular, it is important to provide the bilingual books in this format. For the creation of the *mobi* e-book we use Kindlegen[16]. This utility creates a *mobi* file from the *epub* file. These *mobi* files will be correct displayed in modern readers, but in old ones the links between source and target segments may still appear underlined. To convert the *lip-fra-cat.docbook.epub* into the *lip-fra-cat.mobi* we can run this program in the terminal as follows:

```
kindlegen -c1 -o lip-fra-cat.mobi lip-fra-cat.docbook.epub
```

3.5.3. Conversion to html

We also convert our bilingual e-books to html to allow reading them in any computer or other device with an Internet browser. To convert the bilingual book to html we use the *xsltproc* tool:

```
xsltproc -o lip-fra-cat.html /usr/share/xml/docbook/stylesheet/docbook-
xsl-ns/xhtml-1_1/docbook.xsl prova.docbook lip-fra-cat.docbook
```

It will produce a single *html* page containing all the book. In this *html*, by default, all the links will appear as underlined and in blue. To show the links in black with no underline, we have to add the following line in the head section:

```
<style type="text/css">
a {color: \#000;text-decoration: none}
</style>
```

4 Creation of bilingual dictionaries from free lexical resources

Dictionaries can be of great help for readers in a foreign language, especially when they are integrated in the reading device, allowing a quick search. Some authors point out, however, that the frequent use of dictionaries can prevent the reader from developing guessing skills (Prowse, 2002). Nevertheless, we provide a set of bilingual dictionaries automatically created from free lexical resources, and we let the decision of using them to the readers.

[16]Kindlegen, https://www.amazon.com/gp/feature.html?docId=1000765211

In this section, the process of creation of dictionaries from free lexical resources is explained in a general way. We leave the details on how to use the scripts for the creation of dictionaries for a future paper, as this is not the main issue of this paper.

4.1 Free lexical resources

There are several projects providing free lexical resources that can be used to construct bilingual dictionaries for the e-book readers. It is very important that the resources are distributed with a free licence, allowing reuse and redistribution without legal problems. All the dictionaries we construct can be freely distributed and used. In the following subsections, we present the resources we have used for the creation of the bilingual dictionaries. In table 4 we can see the number of entries of each resource for four language pairs: English-Spanish, English-French, English-Russian and English-Catalan. These figures are approximate, as the resources are growing every day.

eng-spa
eng-fra
eng-rus
eng-cat

	eng-spa	eng-fra	eng-rus	eng-cat
Omegawiki	50,271	37,929	10,737	5,408
Wiktionary	65,660	72,809	101,751	25,286
Wikipedia	919.638	1,229.637	814,122	356,482
Wordnet	57,764	102,671	-	70,624

Table 4: Number of entries of each resource for several language pairs (source: author)

4.1.1. Omegawiki

Omegawiki[17] is a collaborative dictionary. Its aim is to create a dictionary of all words of all languages, including lexical, terminological and ontological information. All the data is available for download as a relational database, and can be easily reused in other projects and applications. The data is available both under the GFDL (*GNU Free Documentation Licence*) and the CC-by licence (*Creative Commons*).

[17] Omegawiki, http://www.omegawiki.org

4.1.2. Wiktionary

Wiktionary[18] is also a collaborative dictionary. This is a sister project of Wikipedia and it's edited using a wiki environment. The content is also under a dual licence: the CC by-sa (*Creative Commons Attribution-ShareAlike 3.0 Unported Licence*) and the GFDL (*GNU Free Documentation Licence*). All the content is downloadable in a wiki-like format and it's not easy to parse to get the relevant data. Several toolkits have been created in order to get structured data from the downloaded files, as JWKTL (*Java-based Wiktionary Library*)[19] (Zesch et al., 2008). The Dbnary project[20] (Sérasset, 2015) provides multilingual lexical data extracted from Wiktionary in an easy to parse format (*Linguistic Linked Open Data*). The use of Dbnary has the advantage of the easy access to the data, but the data in the Wiktionary dumps are updated more regularly.

4.1.3. Wikipedia

Wikipedia[21] is a collaborative multilingual Encyclopedia. It uses a wiki format, as Wiktionary, and the content can be freely downloaded. It is also difficult to parse. The JWPL (Java Wikipedia Library)[22] (Zesch et al., 2008) can be used to extract its content. Dbpedia[23] (Lehmann et al., 2015) is a collaborative project to extract structured information from Wikipedia and make this information available on the Web. The Dbpedia data can be downloaded in an easy to parse format. Here again, the data in the Wikipedia dumps are updated in a more regular basis than the Dbpedia data.

4.1.4. WordNet

WordNet (Fellbaum, 1998) is a lexical database where nouns, verbs, adjectives and adverbs are organized in sets of synonyms called *synsets*. These synsets are connected by semantic relations as hiponymy, antonomy, meronomy, troponomy, etc. The original WordNet was created for English in the Princeton University and now there are wordnets for several languages[24]. In the Open Multilingual WordNet project website[25] (Bond and Kyonghee, 2012) some of these wordnets holding a free licence are published under a common format.

4. 2. Creation of the tab delimited dictionaries

All the data from the lexical resources is stored in a SQLite database. In this way, we are able to access to the data regardless the original format. The database is organized in the

[18] Wiktionary, http://www.wiktionary.org

[19] Java-based Wiktionary Library, https://www.ukp.tu-darmstadt.de/software/jwktl/

[20] Dbnary project, http://kaiko.getalp.org/about-dbnary/

[21] Wikipedia, http://www.wikipedia.org

[22] Java Wikipedia Library, https://www.ukp.tu-darmstadt.de/software/jwpl/

[23] Dbpedia, http://wiki.dbpedia.org/

[24] In the website of the Global WordNet Association http://www.globalwordnet.org a list of existing wordnets is available

[25] Multilingual WordNet project, http://compling.hss.ntu.edu.sg/omw/

following tables:

- *entry*: each entry holds the information about a meaning_id and a word for every source (all lexical resource and wordnets in OMW, included PWN-3.0). It also stores: POS (*Part of Speech*), language and source. Additionally, a word_n field is included to store a normalized word (the word in lower case, without sections in brackets or comma separated). This special field is used with Wikipedia entries.
- *definition*: where the definition for each meaning_id (if available) is stored, along with the language and source.
- *relation*: in this table all the related words for a given meaning_id are stored along with the relation name, language and source.
- *translation*: this table stores all the target language words for a given meaning_id, along with the language and source.

Using this database, tab delimited dictionaries are created with an entry for each lemma-POS combination. Three types of dictionaries are constructed:

- Monolingual dictionaries: the dictionary contains the words in a source language (English in the example) with their POS and definitions in the source language.

```
atmosphere    (n) the mass of air surrounding the Earth
```

- Bilingual dictionaries: the dictionary contains the words in a source language (English in the example) with their POS and the translation to a target language (French in the example).

```
atmosphere    (n) Atmosphère
```

- Combined dictionaries: this is a combination of the previous two types. The dictionary contains the words in a source language with their POS. Then, when available, the definitions in the source language. Lastly, when available, the translations in the target language (French in the example).

```
atmosphere    (n) the mass of air surrounding the Earth. Atmosphère
```

4.3. Morphological expansion of the entries

In most e-book readers, the search in a dictionary is performed searching for exact matches. As words can appear in several morphological forms, and we want to retrieve the correct information for any of these word forms, an algorithm for the morphological expansion of the entries has been developed.

To expand the entries, we use the morphological dictionaries of the Freeling analyzer[26] (Padró and Stanilovsky, 2012). These dictionaries provide word form, lemma

[26] Freeling, http://nlp.lsi.upc.edu/freeling/

and tag for a lot of word forms. In table 5 we can observe the number of entries of the Freeling morphological dictionaries for several languages. Note that than the richer the morphology of a language, the bigger the number of entries.

Language	Entries
Catalan	642,437
English	88,770
French	452,639
Russian	2,544,516
Spanish	669,216

Table 5: Number of entries of the Freeling morphological dictionaries for several languages (source: author)

Let's take as example the following 3 entries from the English-Spanish dictionary:

```
cut       (a) cortado
cut       (n) corte
cut       (v) cortar
```

Here we can observe the information provided from the English dictionary for the word of our example (searching by lemma)

```
cut cut NN
cut cut VB
cut cut VBD
cut cut VBN
cut cut VBP
cuts cut NNS
cuts cut VBZ
cutting cut VBG
```

This information allows us to expand the entries in the dictionary. After the morphological expansion, we have 6 entries.

```
cut      (a) cortado
cut      (n) corte
cut      (v) cortar
cuts     (n) corte
cuts     (v) cortar
cutting  (v) cortar
```

4.4. Merging of all information for each word

After the processes described so far, we have one entry for each word form and POS. As the final formats for the dictionaries need a single entry for each different word, in

this step we merge all the information for each word. So, for example cut can be a verb (cut, cuts, cutting) and a noun (cut, cuts) and an adjective (cut), before the merging step we had the following entries:

In this step, we merge all this information and a single entry for each word form is created:

```
cut (a) cortado; (n) corte; (v) cortar
cuts (n) corte; (v) cortar
cutting (v) cortar
```

Note that we have an entry for each word form, but the corresponding translations are always in the lemma form. Further processing should be necessary to have the target words in all their inflected forms. For target language native users, the lemma form is enough to understand the meaning.

We have created dictionaries for all the languages combinations with published books so far, that is: English-Spanish, French-Spanish, French-Catalan and Russian-Spanish. We have also created some English dictionaries with other target languages with no books published yet. All these dictionaries can be used with any book, so they can be interesting for many users. We plan to publish new dictionaries when books in new language pairs will be published. In table 6 the number of entries of the published bilingual dictionaries are presented. Please, keep in mind that the figures are about word forms, as one entry for each word form is created. This is the reason why the number of entries is so high:

Language pair	Entries	Entries exp.
English - Catalan	86,466	135,695
English - Dutch	74,921	116,105
English - French	133,139	186,818
English - Italian	95,237	147,032
English - Russian	67,059	106,564
English - Spanish	115,425	166,713
French - Catalan	52,462	130,428
French - Spanish	82,495	196,886
Russian - Catalan	30,817	52,514
Russian - Spanish	62,493	117,628

Table 6: Number of entries of the bilingual dictionaries before and after the morphological expansion (source: author)

4.5. Transformation of the dictionaries

Different devices use different dictionary formats. To transform the tab delimited dictionary into the target formats we are using Penelope[27], a set of tools for creating, editing and converting dictionaries. The tool is designed to work with the major dictionary formats for e-book reader devices. Using this tool, we can create dictionaries in the following formats:

- StarDict
- Kobo
- Booken Cybook Odyssey
- XML
- Epub
- Mobi

The *epub* and *mobi* dictionaries are in fact electronic books containing the entries and are suitable for those devices with no automatic search functionality.

5. Multilingual translation memories

As a derived product from the InLéctor project, we are distributing the translation memories created by the alignment of the books. These translation memories can be useful in several situations (Zanettin, 2002):

- For teaching literary translation, as a lot of examples on how to translate a given expression can be found.
- For literary translators. There is not a tradition on the use of Computer Assisted Tools in translation, as a very low degree of fuzzy matches are expected. However, the use of parallel corpora or translation memories can be useful to search how a given expression has been translated before.
- For traductological research on several aspects of translation.

The translation memories are published in the standard Translation Memory Exchange format and can be used with any Computer Assisted Translation Tool. They are also distributed in plain text format, so they can be also used in any corpus processment tool. They are published in the InLéctor Sourceforge page, and we plan to publish them also in Opus Corpus[28] (Tiedemann, 2012).

In table 7 we can observe the number of segments and words in the translation memories for each language pair.

[27] Penelope, https://github.com/pettarin/penelope
[28] Opus Corpus, http://opus.lingfil.uu.se/

Language pair	Segments	L1 words	L2 words
English - Spanish	14,361	247,723	244,846
French - Catalan	4,277	79,208	77,653
French – Spanish	20,588	244,641	232,321
Russian – Spanish	2,248	17,889	18,749

Table 7: Number of segments and words in the translation memories for each language pair (source: author)

6. InLéctor collection

6.1. Published books

We have published 8 titles in the InLéctor collection:

- Arthur Conan Doyle, *The Adventures of Sherlock Holmes* - English-Spanish
- Jane Austen, *Sense and Sensibility* - English-Spanish
- Robert Louis Stevenson, *Strange Case of Dr Jekyll and Mr Hyde* - English-Spanish
- Alexandre Dumas, *Les Trois Mousquetaires* - French-Spanish
- Anatole France, *L'Île des pingouins* - French-Catalan
- Фёдор Достоевский, *Игрок* – Russian-Spanish
- Herman Melville, Mobi Dick, English-Spanish
- Daniel Defoe, Robinson Crusoe, English-Spanish

In all cases, except Anatole France's "L'Île des Pingouins" (French-Catalan), both original text and translations are in the public domain and the texts have been obtained from Wikisource or Project Gutenberg. The two last titles, Moby Dick and Robinson Crusoe have been created in collaboration with Garcitextos[29], a collection of bilingual books created by the Professor Emeritus of the Duke University Miguel Garci-Gomez.

7. Conclusions and future work

We have presented the process of automatic creation of bilingual e-books in the InLéctor collection. All the programs used for the creation of the books hold a free licence and can be used by anyone with no licence costs. A set of Python scripts have been created and are also distributed with a free licence. Using these tools the creation of a bilingual e-book, once we have the source and target text, can be performed in a couple of hours.

[29]Garcitextos, http://torocitydesigns.com/garcitextos/

We have also presented the process of creation of dictionaries from free lexical resources. These dictionaries are published in suitable formats for e-book readers and can be used, not only with the books of the InLéctor collection, but with any electronic book. As the lexical resources used for the creation of the dictionaries hold a free licence, the resulting dictionaries can be also distributed with a free licence.

The aim of the project InLéctor is to promote reading in a foreign language. We are regularly publishing books in the public domain (both source and translation), but the same procedure can be used for publishing any book, provided that the publisher holds the copyright for the original and the translation.

All the books and bilingual dictionaries can be freely downloaded from the InLéctor website [30]. All the algorithms, Python scripts and alignment dictionaries can be downloaded from the SourceForge page of the project[31]. A tutorial on how to create bilingual e-books is also available, so any user can easily create bilingual e-books.

8. References

Bamford, J.; Day, R. R. (1997). "Extensive reading: What is it? Why bother?" *The Language Teacher*, v. 21, n. 5. p. 6-8. <http://jalt-publications.org/tlt/articles/2132-extensive-reading-what-it-why-bother>. [Consulted: January 3, 2017].

Bird, S. (2006). "NLTK: the natural language toolkit". In: *Proceedings of the COLING/ACL on Interactive presentation sessions, Association for Computational Linguistics*, p. 69-72. <https://dl.acm.org/citation.cfm?id=1225403&picked=prox>, <doi 10.3115/1225403.1225421>. [Consulted: April 15, 2017].

Bond, F.; Kyonghee, P. (2012). "A Survey of WordNets and their Licenses". In: *Proceedings of the 6th International Global WordNet Conference, Matsue, Japan*, p. 64-71. <https://research.vu.nl/en/publications/proceedings-of-the-6th-global-wordnet-conference-matsue-japan>. [Consulted: October 11, 2017].

Chin-Neng, C. [et al.] (2013). "The effects of extensive reading via e-books on tertiary level efl students' reading attitude, reading comprehension and vocabulary". *TOJET: The Turkish Online Journal of Educational Technology*, v. 12, n. 2 (April).<http://www.tojet.net/articles/v12i2/12228.pdf> [Consulted: April 15, 2017].

Ernst-Slavit, G.; Mulhern, M. (2003). "Bilingual books: Promoting literacy and biliteracy in the second-language and mainstream classroom". *Reading online*, v. 7, n. 2, p. 1096-1232. <https://ncela.ed.gov/rcd/bibliography/BE022179>. [Consulted: January 3, 2017].

Fellbaum, C. (ed.) (1998). *WordNet: An electronic lexical database.* London: The MIT Press. (Language, speech and communication).

Forcada, M. L. [et al.] (2011). "Apertium: a free/open-source platform for rule-based machine translation". *Machine translation,* v. 25, n. 2, p. 127-144. <doi 10.1007/s10590-011-9090-0>. [Consulted: October 11, 2017].

[30] InLéctor, https://inlector.wordpress.com/

[31] InLéctor available in Sourcefourge at https://sourceforge.net/projects/inlector/

Hafiz, F. M.; Tudor, I. (1989). "Extensive reading and the development of language skills", *ELT journal*, v. 43, n. 1, p. 4-13. <https://doi.org/10.1093/elt/43.1.4>. [Consulted: September 17, 2017].

Nuttall, C. (1996). *Teaching reading skills in a foreign language.* London: Heinemann.

Iribarren, T.; Oliver, A.; Peiró, E. (2016). "Recuperar traduccions inèdites per a internautes: el cas de L'illa dels pingüins, d'Anatole France, en traducció de J. F. Vidal Jové". In: Bacardí, M.; Godayol, P. (eds). *Traducció i franquisme*. Lleida: Punctum. (Visions; 8).

Kiss, T.; Strunk, J. (2006). "Unsupervised multilingual sentence boundary detection". *Computational Linguistics*, v. 32, n. 4 (December), p. 485-525. <doi: 10.1162/coli.2006.32.4.485>. [Consulted: September 18, 2017].

Krashen, S. D. (2004). *The Power of Reading: Insights from the Research: Insights from the Research.* 2nd ed. Santa Barbara, California: ABC-CLIO.

Lehmann, J. [et al.] (2015). "DBpedia: A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia". *Semantic Web Journal*, v. 6, n. 2, p. 167-195. <doi: 10.3233/SW-140134>. [Consulted: September 18, 2017].

Padró, L.; Stanilovsky E. (2012). "FreeLing 3.0: Towards Wider Multilinguality". In: *Proceedings of the Language Resources and Evaluation Conference (LREC 2012), ELRA, Istanbul, Turkey*, p. 2473-2479. <http://www.lrec-conf.org/proceedings/lrec2012/pdf/430_Paper.pdf>. [Consulted: September 18, 2017].

Prowse, P. (2002). "Top ten principles for teaching extensive reading: A response". *Reading in a Foreign Language*, v. 14, n. 2 (October), p. 136-141. <http://nflrc.hawaii.edu/rfl/October2002/day/day.html>. [Consulted: September 17, 2017].

Semingson, P.; Pole, K.; Tommerdahl, J. (2015). "Using Bilingual Books to Enhance Literacy around the World". *European Scientific Journal,* v. 3 (February), p. 132-139. <http://eujournal.org/index.php/esj/article/view/5216/5014>. [Consulted: October 22, 2017].

Sérasset, G. (2015). "DBnary: Wiktionary as a Lemon-based multilingual lexical resource in RDF". *Semantic Web Journal*, v. 6, n. 4, p. 355-361. <http://www.semantic-web-journal.net/system/files/swj648.pdf>, <doi: 10.3233/SW-140147> . [Consulted: October 22, 2017].

Tiedemann, J. (2012). "Parallel Data, Tools and Interfaces in OPUS". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, may,* p. 2214-2218. <http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf>. [Consulted: September 18, 2017].

Varga, D. [et al.] (2007). "Parallel corpora for medium density languages". *Amsterdam Studies in the Theory and History of Linguistic Science Series IV, v. 292*. <https://catalog.ldc.upenn.edu/docs/LDC2008T01/ranlp05.pdf>. [Consulted: October 22, 2017].

Zanettin, F. (2002). "Corpora in Translation Practice". In: Yuste-Rodrigo, E. (ed.). *Language Resources for the Translator Work and Research. LREC 2002 Workshop*

Proceedings, pp. 10-14. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.136.5669&rep=rep1&type=pdf>. [Consulted: September 18, 2017]

Zesch, T.; Müller, C.; Gurevych, I. (2008). "Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary". *LREC*, v. 8, p. 1646-1652. <https://pdfs.semanticscholar.org/065a/29adca32f66c16005de3f48ebb3512c8baf1.pdf>. [Consulted: September 18, 2017].