Multilingual Aligned Corpora From Movie Subtitles

Mathieu Mangeot, Emmanuel Giguet

Multilingual Aligned Corpora From Movie Subtitles

Mathieu Mangeot
Condillac-LISTIC
Université de Savoie
F-73376 Le Bourget du Lac
Mathieu.Mangeot@univ-savoie.fr

Emmanuel Giguet
GREYC
Université de Caen
BP 5186 F-14032 Caen Cedex
Emmanuel.Giguet@info.unicaen.fr

Abstract

This paper describes a methodology for building aligned multilingual corpora form movie subtitles found on the Web. The subtitles have specific formats and encodings. In a first step, we convert them to our multilingual subtitle format based on XML. In a second step, we align the subtitle sentences with the time used to display them on the screen. We implemented the tool Jimaku in order to semi-automatically perform these steps. The last step consists in aligning the sentences at the sub-sentence level and to index the corpus for contextual lookup. For this step, we use the WIMS platform, result of previous research on text collections management.

1 Introduction

Nowadays, lots of Natural Language Processing works use multilingual aligned corpora. For example, they can help to build a terminological database or enrich a translation memory. In this paper, we tackle the building of such a resource with an original approach: the use of movie subtitles. As mentioned by (Simões, 2004), we use a positive side effect of DivX movie piracy on the Internet. It allows us to find an impressive amount of freely available movie subtitles free of rights in many languages.

With this method, one can build quickly and with low cost multilingual aligned corpora by relying on a specific criterion of this data: the temporal indexing of subtitle lines. Furthermore, it allows one to align language pairs for which the opportunities to find other aligned corpora are very weak. One of the first planned use of these corpora is the building of multilingual examples for the Papillon dictionary with a pivot structure (Mangeot et al., 2004).

We first present some sample of corpora in their original format as well as in our target XML representation then we detail the temporal alignment procedure. The next sections are dedicated to the presentation of the tools: Jimaku for the conversion and temporal alignment and WIMS platform (Giguet, 2005a) for aligning the sub-sentences and putting them into context with a concordance program.

2 Presentation of the corpora

2.1 The source: Websites for exchanging subtitles

The growing number of Websites for exchanging movie subtitles followed naturally the increasing rate of DivX downloads. These sites are all based on the same principle: they offer a search in a subtitle database. The query of a subtitle is usually based on two criteria: the movie title and the language subtitle. Usually, the users have to sign in for downloading the subtitle matching their request. They can also upload missing subtitles on the server.

The subtitle websites are most often specialized. Some by movie genre and most of them by language or language group. The majority of the subtitles found on these sites are written by groups of enthusiastic amateurs that begin their work as soon as the movie is available on the Internet. They are available freely but a more detailed study should be driven to check the legal aspects, regarding author copyrights and intellectual property.

2.2 The subtitle formats

Lets present first three samples of subtitles of the movie "Ace Ventura: Pet Detective" in their original format (Figures 1, 2 and 3) but after conversion in UTF-8.

The format of Figure 1 is the following. The time when the subtitle line display is started is on the first line. The time when the subtitle line display is ended is on the second line between brackets. The third line as well as the following lines until a new line between brackets represent the subtitle text that has to be displayed. Then, the following subtitle begins with a new line between brackets. If the text is on several lines, it will be displayed on several lines on the screen. This format is the SubViewer version2. The original encoding of the file is Shift-Jis (Japanese encoding).

The SubRip format seen on Figure 2 is different. The subtitle line number is on the first line and begins by 1 (the first subtitle). The starting and ending time of the display of the subtitle are on the second line, separated by an arrow. The third line, as well as the following lines until an empty one represent the subtitle text that has to be displayed. THe following subtitle begins with a new line with its number only.

the format of Figure 3 is the MicroDVD one. Every-

[00:01:58.920]
[00:02:01.200]
ダウンタウンまでかっ飛ばすぜ
[00:02:09.720]
[00:02:13.080]
黙れ この馬鹿犬!
[00:02:13.720]
[00:02:14.800]
何の用だ?
[00:02:15.040]
[00:02:19.720]
こんばんわ HDSの者ですが
お届け物です
[00:02:22.760]
[00:02:24.120]
壊れてるみたいだぞ

Figure 1: Sample of Japanese subtitle of the movie "Ace Ventura : Pet Detective" in SubViewer 2 format

3
00:01:58,932 −− > 00:02:01,207
Πάμε στην πόλη.

4
00:02:15,052 −− > 00:02:19,728
Απ' την HDS, κύριε. Τι κάνετε;
΄Εχω ένα πακέτο για σας.

5
00:02:22,772 −− > 00:02:24,125
Ακούγεται σπασμένο.

Figure 2: Sample of Greek subtitle of the movie "Ace Ventura : Pet Detective" in SubRip format

thing is on the same line: he starting and ending time of the display of the subtitle are rounded curly brackets and the text follows. The line breaks are represented by the special character "|". The next line is the next subtitle.

{1600}{1670} Dobra obramba!
{2983}{3077} Gremo proti srediču mesta!
{3353}{3380} Kaj hoče?
{3386}{3540} HDS, gospod. Kako ste kaj to popoldne?
Ureduček potem. Paket imam za vas.
{3579}{3613} Slii se zlomljeno.

Figure 3: Sample of Slovene subtitle of the movie "Ace Ventura : Pet Detective" in MicroDVD format

3 Standardization of the formats: XML-Jimaku

3.1 A Large Variety of Formats

To our knowledge, there is no standard for encoding movie subtitles. THus, there is a large variety of formats. Every subtitle writing tool propose its format: SubViewer 1 and 2 (extension .sub), SubStation Alpha (.sst), SubRip (.srt), SAMI (.smi), etc. Furthermore, some movie viewers also propose their own format (Quicktime for example).

In order to use these subtitles in a multilingual alignment perspective, we chose to convert them into a common format. We thus had to precise the data that need to be standardized. We particularly focused on temporal indexing, subtitles rendering and characters encodings :

- **Temporal indexing** the subtitles are usually indexed on the time in milliseconds from the beginning of the movie video track. But some formats like MicroDVD index the subtitles with the video frame number. This information has to be taken into account because we need to convert them in order to align them on the time with the others.
- **Rendering** Most of the formats do not add rendering information on the subtitle text that has to be displayed. In some formats, the line breaks are represented by a special character ("|" for MicroDVD). Some formats propose rendering for some parts of text, mainly bold and italics. THese informatoins are usually represented by SGML-like tags (<b> and <i>).
- **Character encodings** Despite the creation of Unicode in 1992, we did not find any format that uses by default an encoding of the Unicode table like UTF-8. Furthermore, there is no link between the format of a subtitle and its encoding. One must note the file encoding when downloading it from a subtitle Website if this type of information is available, or use an encoding guesser.

3.2 The XML-Jimaku Format

We chose a common format named *Jimaku*, based on XML and Unicode. It can represent all the existing encodings. The root element of the format is `jimaku`. It contains the elements `titles` (for storing the titles in their multilingual diversity), `style` (for linking rendering information), `head` (for the usual meta-information) and `subtitles` (for storing subtitles with temporal indexing).

The `subtitles` element contains a list of `st` elements. The element `st` represents a subtitle. It contains the elements `start` and `end` for the temporal indexing of the subtitle and a list of `text` elements with, for each an attribute `xml:lang` that indicates the language of the subtitle at the format ISO-639-1. The DTD of the Jimaku format is available at the

following address: http://clips.imag.fr/geta/services/dml/jimaku.dtd

3.3 A Jimaku Subtitle Example

The Figure 4 is a sample of a multilingual subtitle of the movie "Ace Ventura: Pet Detective" in Jimaku format. It has been obtained from the conversion and alignment of previous subtitles.

```
<jimaku version="1.0" ov="en"
languages="de,el,en,ja,nl,sl,sv">
<titles>
 <title xml:lang="en">Ace Ventura - Pet
Detective</title>
</titles>
<subtitles>
[...]
<st n="10">
<start>00:02:26.682</start>
<end>00:02:27.722</end>
<txt xml:lang="de">Was?</txt>
<txt xml:lang="en">What do you
want?</txt>
<txt xml:lang="ja">何の用だ?</txt>
<txt xml:lang="sl">Kaj hoče?</txt>
</st>
<st n="11">
<start>00:02:28.002</start>
<end>00:02:32.682</end>
<txt xml:lang="de">HDS, Sir.  Wie geht
es lhnen?<br/>Also gut.  Ein Paket fr
Sie.</txt>
<txt xml:lang="el">Απ' την HDS, κύριε.  Τι
κάνετε;<br/>Έχω ένα πακέτο για σας.</txt>
<txt xml:lang="en">HDS, sir.  How are
you this afternoon?<br/>Alrighty.  I
have a package for you.</txt>
<txt xml:lang="ja">こんばんわ
HDSの者ですが<br/>
お届け物です</txt>
<txt xml:lang="nl">- Wat moet je
?<br/>- HDS. Alles goed ?  Mooi
zo.</txt>
<txt xml:lang="sl">HDS, gospod.  Kako
ste kaj to popoldne?  Ureduek potem.
Paket imam za vas.</txt>
<txt xml:lang="sv">HDS, sir.  Hur str
det till i dag?<br/>Jag har ett paket t
er.</txt>
</st>
[...]
</subtitles>
```

Figure 4: Sample of multilingual subtitle of the movie "Ace Ventura : Pet Detective" in XML-Jimaku format

4 Subtitle Alignment

4.1 Manual Alignment

Even if they are created for the same movie, the different subtitles have very often a time shift between each others. There are several reasons for that: the video track used to create the subtitle is not always the same, the videos do not start at the same time, etc...

Furthermore, the conversion between formats and particularly between frame and time indexed formats can generate time shifts because of the frames per second rate of the video.

Before the subtitle alignments, it is thus necessary to reset the subtitles alignment in order to set them on the same video track (hypothetical). The easiest method consists in aligning a subtitle line at the beginning of each file and another one at the end. Then, the computation is propagated on all other subtitle lines.

4.2 Automatic Alignment on the Time

The alignment of several subtitles on the time measured between the beginning of the movie and the display of a subtitle on the screen is the core step of our methodology.

Some subtitles are built from copies of other existing subtitles. Every line is translated in a new language and the time is not changed. These subtitles do not cause any particular problem because they are already aligned. We just gather the lines that have the same time.

The previous case do not occurs very often. Most of the time, every subtitle file is built independently fro the others directly from the original video track. As a result, the times are never identical at the milli-second level.

A third case also occurs frequently: when the subtitle lines are indexed on the frame numbers in the original format, they must be first converted into time indexing. Here also, the time cannot perfectly match.

In order to handle the last two cases, the subtitle line alignments is done on the time with a variation (delta t). We fixed empirically this delta t at 500 milliseconds. Above, the risk to align un-corresponding subtitle lines thus that are not translations the ones from the others is too important.

5 Subtitle Management with Jimaku Tool

5.1 Presentation of the Tool

The Jimaku tool (*subtitle* in Japanese) has been designed and implemented by Mathieu Mangeot and David Thevenin on a common original idea. the first version was launched in 2003. for the moment, this tool is not downloadable. Nevertheless, interested people can contact us.

The first aim of this tool was to be able to reset the alignment of a subtitle on its video track. The format

and encoding conversion functionalities where useful from the beginning.

5.2 Format Management

Because of the large amount of formats, we had to adapt the structure of our tool in order to facilitate the management of a new format. For that purpose, a programmer just has to implement an interface (abstract java class) that can read from the input this new format and convert them into the Jimaku internal representation format and also to write to the output in this format. The formats already handled are the following:

- **Jimaku** multilingual format, based on XML, indexed on the time, extension *.jmk*
- **MicroDVD** monolingual format, indexed on the frame, extension *.sub*
- **QuickTime** monolingual format, indexed on the time, extension *.txt*
- **SAMI** multilingual format, based on HTML, indexed on the time, extension *.smi*
- **SubRip** monolingual format, indexed on the time, extension *.srt*
- **SubViewer 1 and 2** monolingual format, indexed on the time, extension *.sub*
- **SubStation Alpha** monolingual format, indexed on the time, extension *.ssa*

When a subtitle file is opened into Jimaku, the format is automatically detected and the subtitles are converted into the internal representation format. The subtitles can also be edited and the alignment modified. The "Save" command save the changes in the original format. The "Export" command allows the user to convert the subtitle file into another format and write a result file in output.

5.3 Encoding Detection

The encoding and file languages are almost never specified in the subtitle file contents. If the information has not been obtained elsewhere, the user has to test several encodings in order to find the correct original encoding of the file. The Jimaku tool provides a function to change the encoding on the fly. Note: the Jimaku format is always encoded in UTF-8, (furthermore, XML specifies the encoding of the files) thus there is no need to specify the encoding.

5.4 Temporal Resetting Functions

As explained previously, it is often necessary to reset the alignment of a subtitle with the corresponding video track. the temporal resetting functions of Jimaku are the following: shifting all the subtitles by a period t, "stretching" the subtitles and manual resetting. The shift can be done either in milliseconds, either by indicating the new time of a subtitle line, or in frame numbers. The stretch can be done either from 24 to 25 frames per second, or the contrary. The manual resetting is done by indicating the time of a subtitle line at the beginning of the movie and another one at the end.

5.5 Temporal Alignment Functions

The temporal alignment of two files is launched when an external file is imported into an already opened one. The latter must be in Jimaku format because the result of the import is a multilingual subtitle file.

During the import, the user has to specify the location of the file to import as well as its encoding and language. Then, the manual alignment interface appears (such as Figure 6). the subtitles must be aligned on two lines preferably far the one from the other. This is why the interface shows a window with the first subtitle lines and another window with the last ones. According to our experience, it is easier to align the subtitle lines containing proper nouns because in an unknown language it is easier to identify them than other words. The figure 6 shows an alignment first on "HDS" and then on "Mr. Ace Ventura".

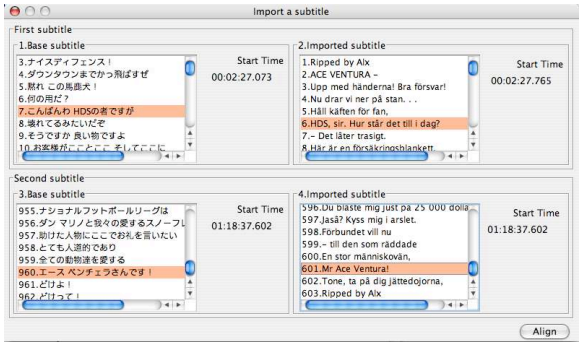
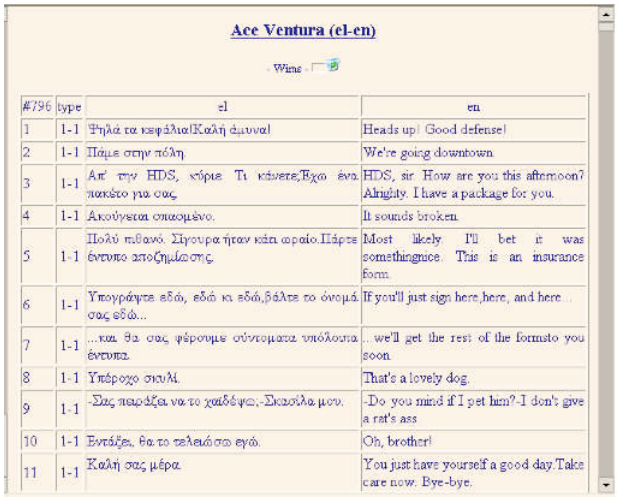


Figure 5: Manual subtitle alignment with Jimaku

6 Subtitle Management in WIMS Platform

The multilingual subtitle files produced with Jimaku are aligned at the subtitle line level thanks to the temporal criterion. They are then imported into the WIMS platform. This platform manages the presentation of the documents, their diffusion, and also offers the opportunity to apply several linguistic analysis. WMS integrates, among others, a sub-sentence alignment module that we adapted in order to analyze the Jimaku files. The XML structure of Jimaku files is close to the TMX standard already handled by our alignment module. We wrote an XSLT stylesheet in order to convert the Jimaku files into the TMX format.

The sub-sentence alignment method included in WIMS has been already published in several papers (Giguet, 2005b; Giguet and Apidianaki, 2005). We will thus present it only briefly in this paper. Lets just



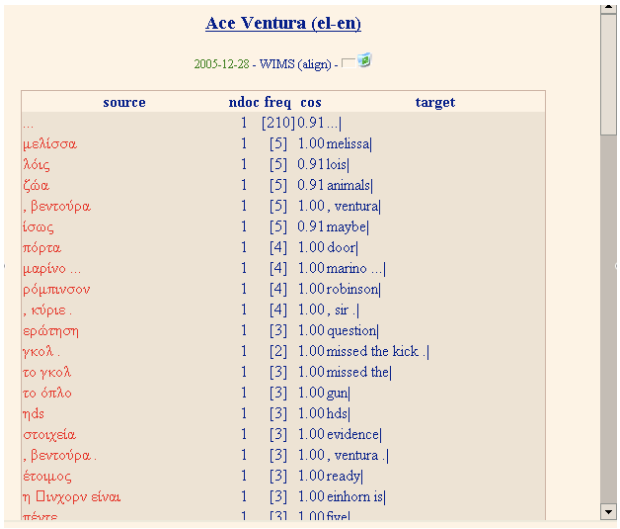
Ace Ventura (el-en)

. Wims .

#796	type	el	en
1	1-1	Ψηλά τα κεφάλια!Καλή άμυνα!	Heads up! Good defense!
2	1-1	Πάμε στην πόλη.	We're going downtown.
3	1-1	Απ' την HDS, κύριε. Τι κάνετε;Έχω ένα πακέτο για σας.	HDS, sir. How are you this afternoon? Alrighty. I have a package for you.
4	1-1	Ακούγεται σπασμένο.	It sounds broken.
5	1-1	Πολύ πιθανό. Σίγουρα ήταν κάτι ασφαλίσιμο.Πάρτε έντυπο αποζημίωσης.	Most likely. I'll bet it was somethingnuce. This is an insurance form.
6	1-1	Υπογράψτε εδώ, εδώ κι εδώ,βάλτε το όνομά σας εδώ...	If you'll just sign here,here, and here...
7	1-1	...και θα σας φέρουμε σύντομα υπόλοιπα έντυπα.	...we'll get the rest of the formsto you soon.
8	1-1	Υπέροχο σκυλί.	That's a lovely dog.
9	1-1	-Σας πειράζει να το χαϊδέψω;-Σκασίλα μου.	-Do you mind if I pet him?-I don't give a rat's ass
10	1-1	Εντάξει, θα το σελαιώσω εγώ.	Oh, brother!
11	1-1	Καλή σας μέρα.	You just have yourself a good day.Take care now. Bye-bye.

Figure 6: Subtitle presentation in WIMS

stress the fact that the method is endogenous: it does not use any other resource apart from the available corpus. This approach can handle transparently poorly represented languages or specialized fields texts (medical for example).



Ace Ventura (el-en)

2005-12-28 - WIMS (align) -

source	ndoc	freq	cos	target
...	1	[210]	0.91	...|
μελίσσα	1	[5]	1.00	melissa|
λόις	1	[5]	0.91	lois|
ζώα	1	[5]	0.91	animals|
, βεντούρα	1	[5]	1.00	, ventura|
ίσως	1	[5]	0.91	maybe|
πόρτα	1	[4]	1.00	door|
μαρίνο ...	1	[4]	1.00	marino ...|
ρόμπινσον	1	[4]	1.00	robinson|
, κύριε .	1	[4]	1.00	, sir .|
ερώτηση	1	[3]	1.00	question|
γκολ .	1	[2]	1.00	missed the kick .|
το γκολ	1	[3]	1.00	missed the|
το όπλο	1	[3]	1.00	gun|
ηds	1	[3]	1.00	hds|
στοιχεία	1	[3]	1.00	evidence|
, βεντούρα .	1	[3]	1.00	, ventura .|
έτοιμος	1	[3]	1.00	ready|
η Πινχορν είναι	1	[3]	1.00	einhorn is|
πέντε	1	[3]	1.00	five|

Figure 7: Subtitle alignment in WIMS

7 Conclusion

We presented a low cost methodology for quickly building multilingual corpora covering an important amount of languages. The major disadvantage of these corpora is that they are limited to the movie subtitles. Nevertheless, we think that they can be very useful as we plan to demonstrate with the building of multilingual examples for Papillon dictionary.

Our methodology can be improved mainly at the subtitle line alignment. The next planned step is to automatically calculate the delta t that would optimize the subtitle alignment with a maximum precision and minimum recall. The next step consists in improving again the Jimaku original alignment technique with the WIMS textual alignment ones.

References

Emmanuel Giguet and Marianna Apidianaki. 2005. Alignement d'unités textuelles de taille variable. In *Journée Internationale de la Linguistique de Corpus*, Lorient, France, septembre.

Emmanuel Giguet. 2005a. Modélisation de l'activité expérimentale du chercheur en traitement des langues sur corpus multilingues. In *Journée « Articuler les traitements sur corpus », organisée par Benoît Habert, Serge Heiden et André Salem*, Paris, France, 12 février.

Emmanuel Giguet. 2005b. Multi-grained alignment of parallel texts with endogenous resources. In *Recent Advances in Natural Language Processing (RANLP) International Workshop "New Trends in Machine Translations"*, Borovets, Bulgaria, 24 September.

Mathieu Mangeot, Gilles Sérasset, and Mathieu Lafourcade. 2004. Construction collaborative d'une base lexicale multilingue. *Traitement Automatique des Langues*, 44(2):151--176, February.

Alberto Simões. 2004. Parallel corpora word alignment and applications. Technical report, Escola de Engenharia - Universidade do Minho, Minho, Portugal.