A Multimedia Parallel Corpus of English-Galician Film Subtitling

Patricia Sotelo Dios[1] **and Xavier Gómez Guinovart**[2]

1 **University of Vigo, Galicia, Spain**
psotelod@uvigo.es
2 **University of Vigo, Galicia, Spain**
xgg@uvigo.es

Abstract

In this paper, we present an ongoing research project focused on the building, processing and exploitation of a multimedia parallel corpus of English-Galician film subtitling, showing the TMX-based XML specification designed to encode both audiovisual features and translation alignments in the corpus, and the solutions adopted for making the data available over the web in multimedia format.

1998 ACM Subject Classification H.3.1 Content Analysis and Indexing

Keywords and phrases corpora, multimedia, translation, subtitling, XML

Digital Object Identifier 10.4230/OASIcs.SLATE.2012.255

1 Introduction

The CLUVI Corpus is an open collection of parallel text corpora that covers specific areas of the contemporary Galician language. With over 23 million words, the CLUVI Corpus comprises six main parallel corpora belonging to five specialised registers or domains (fiction, computing, popular science, law and administration) and involving five different language combinations (Galician-Spanish bilingual translation, English-Galician bilingual translation, French-Galician bilingual translation, English-Galician-French-Spanish tetralingual translation and Spanish-Galician-Catalan-Basque tetralingual translation). Among the various applications of this corpus, it has been used mainly for lexical extraction in terminology and translation [8, 14, 10, 11, 7].

The format chosen for storing the aligned parallel texts is an adaptation of the TMX format [9], as this is the XML encoding standard for translation memories and parallel corpora, regardless of the application used. A translation memory is a database that collects and records source text segments and their corresponding translated versions with the purpose of being reused for further translations via a computer-aided translation system. Albeit with some differences, an aligned parallel corpus is equivalent to a translation memory. In fact, the last few years have seen an increasing number of TMX-encoded aligned parallel corpora, which offer the additional advantage that they can be used as translation memories for feeding computer-aided translation programs (as proposed in [15]).

In this paper, we present the methodology developed by the SLI (Computational Linguistics Group of the University of Vigo) for building and processing the Veiga Corpus, a multimedia extension of the CLUVI featuring English-Galician cinematographic parallel texts, which is currently underway. In the following section we describe the data and briefly discuss the nature of the corpus. Section 3 deals with the actual construction of the corpus, including annotation and the two-layered segmentation and alignment processes. Section 4

1st Symposium on Languages, Applications and Technologies (SLATE'12).
Editors: Alberto Simões, Ricardo Queirós, Daniela da Cruz; pp. 255–266

OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

arises some discussion about the potential use of such a corpus as a tool for researchers, teachers and subtitling practitioners. And in the last section, we draw some conclusions, point to certain challenges and outline possible future directions in terms of corpus development and research.

2 The Veiga Corpus

The Veiga Corpus is a small (approximately 300,000 words) but ever-growing English-Galician corpus consisting of 24 American, British, and Australian English-language films subtitled in both English (intralingual subtitling) and Galician (interlingual subtitling) for DVD, cinema and Internet distribution. Developed under the broader framework of the CLUVI Corpus, the Veiga was born as a text-only corpus of subtitles. It was not until very recently that we decided to make it multimedia, as soon as we found the appropriate tools to process the data and to make it accessible to the public in what we considered to be an appropriate way. The Veiga multimedia corpus of subtitles is already available for public consultation at `http://sli.uvigo.es/CLUVI/vmm_en.html`. However, it should be noted that only 13 of the 24 films are available in multimedia format at this writing.

Unlike the other CLUVI corpora, strictly speaking the Veiga Corpus cannot hold the title of parallel, nor can it be given the label of comparable, in accordance with how these concepts are traditionally understood and defined in the literature. The term 'parallel corpora' usually refers to 'original, source-language texts in language A and their translated version in language B' [1]. In contrast, a comparable corpus can be defined as a document collection composed of two or more disjoint subsets of documents, each written in a different language but dealing with the same topic.

Thus, the Veiga Corpus inhabits a certain intermediate land between a parallel and a comparable corpus. On the one hand, the Galician subtitles cannot be deemed to stand for translations of the English subtitles, although it could also be the case that subtitlers used the English file (if available) when translating into Galician. On the other hand, these two subsets share more than their semantic content: they both could be considered versions of the same original audiovisual text. Hence, we could say that the relationship among the original English subtitles and the Galician subtitles is triangular shaped.

The real, strict parallelism would be that occurring between the original text and each of the two subsets of English and Galician subtitles. The English set would correspond to a very particular type of transcription, which is known as intralingual subtitling, and the Galician set would embody an also very special modality of translation, which is given the name of interlingual subtitling. And yet, a parallel relation is very likely to come into play between the two sets of subtitles as well –a peculiar kind of 'cotranslation'–, inasmuch as they both are 'sub-products' of the same original text. In sum, a double unidirectional parallel may be established between the original audiovisual text and the subtitles, and a bidirectional correlation is also expected to exist between the subtitles themselves, as shown in Figure 1.

Obviously, a text-only corpus of subtitles would not allow for any kind of parallel observance in terms of source vis-a-vis translated text. By giving users access to the original audiovisual product some comparisons and parallelisms can be made, providing them with the opportunity to explore the manifold dimensions of subtitles, as for example phenomena related to the semiotics of interlingual and intralingual subtitling.

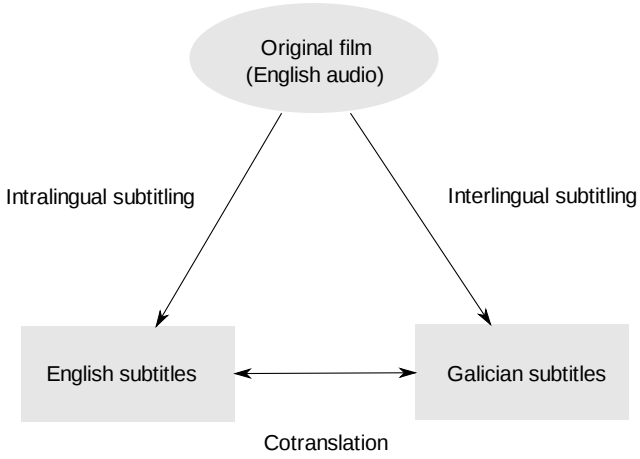

Figure 1 The subtitling triangle.

3 Tagging the Subtitling Triangle

As mentioned before, the Veiga Corpus is hosted at the CLUVI Corpus collection. The CLUVI Corpus functions as a repository of parallel subcorpora of different sizes and thematic fields, all of which undergo identical compiling and processing routines, and can be similarly accessed from one single search interface. Nonetheless, the Veiga Corpus requires further processing in comparison to the other CLUVI subcorpora. Besides annotating stylistic aspects of translation such as omissions, additions and reordering of translation units, all the subtitles include both the in-cue and out-cue time and the line break indicator, allowing users to examine aspects which are inherent to the subtitling practice, e.g. time and space constraints, segmentation, and condensation, among other particularities. In addition to this, the multimedia version of the Veiga incorporates a bonus feature: it enables users to stream the video clips corresponding to the bilingual pairs found in the search results, thus giving them access to the (co-)text in its original, multi-semiotic form. This means that wherever there is a result that matches the query in text format, the search interface shows a link to the corresponding video clips subtitled in each of the two languages involved (English and Galician). All the above mentioned aspects of the Veiga Corpus are annotated according to the TMX-based XML CLUVI specification for parallel corpora, which is summarized in Listing 1.

3.1 Tagging Subtitles at Textual Level

The basic segmentation unit for the alignment of the CLUVI bitexts is the orthographic sentence of the source text. Therefore, the correspondence between source and target text will always be of the 1:n type. The most frequent case is to have one sentence of the source text that corresponds with one sentence of the translation (1:1). Nevertheless, there are instances in which a source sentence is not translated (1:0), or in which a source sentence corresponds with half a sentence (1:1/2) or with two sentences of the translation (1:2), or even in which a sentence of the translation does not correspond with any source sentence (0:1). Moreover, translating frequently implies rearranging and relocating sentences and parts of sentences, in such a manner that these sentences or segments get moved to a different position in the translated text. These elements are reordered in the target section of the CLUVI parallel corpora in order to match the 1:n alignment criterion that preserves the

Listing 1 TMX-based CLUVI specification.

```
<!-- CLUVI_TMX DTD -->
<!ELEMENT cluvi_tmx (header, body) >
<!ATTLIST cluvi_tmx
       version CDATA #REQUIRED >
<!ELEMENT header (#PCDATA)>
<!ELEMENT body (tu*) >
<!ELEMENT tu (tuv+) >
<!ELEMENT tuv (seg) >
<!ATTLIST tuv
       xml:lang CDATA #REQUIRED>
<!ELEMENT seg (#PCDATA | s | l | hi | ph)*>
<!ELEMENT hi (#PCDATA | l)*>
<!ATTLIST hi
       type CDATA #IMPLIED
       x CDATA #IMPLIED>
<!ELEMENT ph EMPTY>
<!ATTLIST ph
       x CDATA #IMPLIED>
<!ELEMENT s EMPTY>
<!ATTLIST s
       n CDATA #IMPLIED
       d CDATA #IMPLIED
       a CDATA #IMPLIED>
<!ELEMENT l EMPTY>
```

integrity and the disposition of the translation units of the source text. This criterion is crucial when applied to the processing of multilingual corpora, where source sentences must provide for the establishment of correspondences among equivalent sentences in various languages.

The TMX specification does not consider the encoding of these translation phenomena, it has been designed for storing and exchanging translation memories and not for representing equivalent segments in parallel corpora. The TMX-based CLUVI encoding system uses an adapted version of some of the tags which are part of the TMX 1.4 specification [13] in order to represent the not-1:1 correspondences and reorderings encoded in the CLUVI parallel corpora. The aspects of translation encoded in the CLUVI corpora can be described as either omission, addition or reordering, and will be tagged using an adapted version of TMX 1.4 content elements <hi> and <ph>.

An omission occurs when an item of the source text does not correspond with any item of the target text, that is, when a sentence or part of a sentence is not translated. Omissions in the CLUVI parallel corpora are encoded by means of the <hi> element. According to the TMX 1.4 specification, the <hi> (or highlight) element 'delimits a section of text that has special meaning, such as a terminological unit, a proper name, an item that should not be modified, etc.' [13]. In the TMX-based CLUVI encoding, the <hi> element marks the piece in the source text that is omitted in the target text. This use of the <hi> tag is noted by means of the type attribute with the "supr" value. For instance, the non-translation of the English source sentence 'Yeah, okay, I'll be there' in the alignment of Wim Wenders' film

Listing 2 Example of omission in the Veiga Corpus.

```
<tu>
  <tuv xml:lang="en"><seg><s n="19" d="00:08:30,411" a="00:08:31,924">Oh
      okay, all right.</seg></tuv>
  <tuv xml:lang="gl"><seg><s n="24" d="00:08:31,371" a="00:08:32,599">
      ¡Vale!</seg></tuv>
</tu>
<tu>
  <tuv xml:lang="en"><seg><s n="20" d="00:08:32,011" a="00:08:36,084"><hi
      type="supr">Yeah, okay, I'll be there.</hi></seg></tuv>
  <tuv xml:lang="gl"><seg>[[---]]</seg></tuv>
</tu>
<tu>
  <tuv xml:lang="en"><seg><l/>I'll get there as fast as I can.</seg></tuv>
  <tuv xml:lang="gl"><seg><s n="25" d="00:08:34,411" a="00:08:37,130">Irei
       o máis axiña que poida.</seg></tuv>
</tu>
<tu>
```

Paris-Texas –included in the Veiga Corpus– would be encoded as shown in Listing 2.

On the other hand, the translation technique known as addition involves the insertion of elements in the target text that have no correspondence in the source text. Addition is also encoded in the CLUVI by means of the <hi> element, which highlights the inserted unit in the target text. This use of the <hi> tag is indicated by means of the type attribute with the "incl" value. The added text joins the translation unit into which it is inserted. If the new element is a sentence (or a sequence of sentences), it joins either the preceding or the following translation unit, depending on its context, thus respecting the 1:n alignment criterion, as shown in the example of Listing 3 excerpted from this same film.

Reordering in translation implies moving sentences or parts of sentences from their original position in the source text to a new location in the translated text. These displaced elements are reordered in the target section of the CLUVI parallel corpora to fit with the 1:n alignment criterion that preserves the integrity and the order of the translation units of the source text. Reordering in the CLUVI is encoded by means of a combination of the <hi> element and the <ph> element. The phrase or sentence that is being moved is tagged with the <hi> element, whose type attribute has the value "reord", and an x attribute with a numeric value that acts as an unambiguous index. In addition, a <ph> element in the translated text indicates the original location of the item being moved. According to the TMX 1.4 specification, the <ph> (or placeholder) element is used 'to delimit a sequence of native standalone codes in the segment. Standalone codes are codes that are not opening or closing of a pair, for example empty elements in XML' [13]. In the TMX-based CLUVI encoding, the adapted <ph> element marks the departure point of the moved text block, and the relationship between this piece of text and its place of origin is encoded in the <ph> element by means of an x attribute that has the same value as the index encoded in the corresponding <hi> tag of the segment being moved. Given the need for synchrony between the subtitles and the audiovisual narrative that is distinctive to the practice of subtitling, reorderings are very rare in the Veiga Corpus. The example in Listing 4 –gathered also from the above mentioned film– shows how reorderings are encoded.

Listing 3 Example of insertion in the Veiga Corpus.

```
<tu>
  <tuv xml:lang="en"><seg><s n="89" d="00:18:25,251" a="00:18:27,242">O.K.,
       I'll be right back.</seg></tuv>
  <tuv xml:lang="gl"><seg><s n="100" d="00:18:26,011" a="00:18:27,808">
      Volvo de contado.</seg></tuv>
</tu>
<tu>
  <tuv xml:lang="en"><seg>[[---]]</seg></tuv>
  <tuv xml:lang="gl"><seg><s n="101" d="00:19:01,051" a="00:19:05,010"><hi
       type="incl">CALZADOS<l/>FERRAXERÍA</hi></seg></tuv>
</tu>
<tu>
  <tuv xml:lang="en"><seg><s n="90" d="00:20:13,571" a="00:20:14,845">Damn
       it.</seg></tuv>
  <tuv xml:lang="gl"><seg><s n="102" d="00:20:14,451" a="00:20:15,486">
      ¡Merda!</seg></tuv>
</tu>
<tu>
```

Furthermore, the tagging of the Veiga Corpus includes annotation of line breaks in subtitles. Line breaks within subtitles are encoded in the Veiga with the <l/> tag, an element added to the TMX 1.4 specification to allow for the examination of aspects which are relevant to subtitling, such as typographical conventions and space constraints.

The Veiga texts were aligned using the free alignment software Trans Suite 2000 Align, which performs automatic segmentation and alignment of both source and target texts. This tool operates at the sentence level, meaning that whenever the system detects a punctuation mark in either language, a new segment is identified and created. Considering that suspension dots in subtitling are also used to indicate that the sentence is not finished and will continue in the next subtitle, and that each subtitle timecode is enclosed in square brackets –which are often erroneously recognized by the software as sentence boundaries–, some manual checking and editing of the automated segmentation and alignment needs to be done. Besides, as mentioned earlier, we must comply with the 1:n criterion, which involves segment merging and splitting, mainly in the target side.

3.2 Tagging at the Textual/Audiovisual Interface Level

Tagging the Veiga Corpus at the textual/audiovisual interface level implies, on one hand, tagging the correspondences between the English subtitles stored as XML textual data in the TMX-based CLUVI encoding and the equivalent segment of the original English-language film with English subtitles, and, on the other hand, tagging the correspondences between the Galician subtitles stored as XML textual data and the equivalent segment of the original English-language film with Galician subtitles. In order to be able to establish these textual/audiovisual correspondences, all of the Veiga English-language films have been cut into video clips, each one corresponding to a subtitle. A first step is to check if the subtitles are in sync with the movie. In some cases, mostly when the subtitle file and the movie come from different sources, we need to edit the subtitles (using the freeware tool

Listing 4 Example of reordering in the Veiga Corpus.

```
<tu>
  <tuv xml:lang="en"><seg><l/>Everybody, everybody'll see me.</seg></tuv>
  <tuv xml:lang="gl"><seg><hi type="reord" x="2">-Vanme ver todos.</hi></
      seg></tuv>
</tu>
<tu>
  <tuv xml:lang="en"><seg><s n="354" d="00:47:21,131" a="00:47:24,328"><hi
       type="supr">No, Travis, I insist.</hi></seg></tuv>
  <tuv xml:lang="gl"><seg>[[---]] </seg></tuv>
</tu>
<tu>
  <tuv xml:lang="en"><seg><l/>He'll wait for you out in front.</seg></tuv>
  <tuv xml:lang="gl"><seg><s n="381" d="00:47:22,051" a="00:47:25,009">-
      Esperarate na porta.<ph x="2"/></seg></tuv>
</tu>
```

Subtitle Workshop) and add a time delay (forward or backward) so that their speed matches that of the video. Secondly, we embed the subtitles in the two languages in the original film with a free, open-source video editing tool called VirtualDubMod. And finally, we edit each film subtitled both in English and in Galician and segment it into subtitles.

Therefore, we come up with two subsets of subtitled video clips, one in English and the other in Galician, each made up of as many videos as subtitles has the corresponding film. Moreover, given that a high number of subtitles are not long enough to be played and watched properly (they are only one or two seconds long), each individual clip/subtitle is allotted ten extra seconds –five seconds before the subtitle shows up, and five seconds after it fades out–, thus providing the viewer with some context. Needless to say, this segmentation process is very monotonous and time consuming, which is a major hurdle that could be overcome if we found a freeware video editing tool offering customary, automatic batch splitting features. Now, once we get two sets of subtitled clips for each film, we link them to their corresponding text in the bitextual TMX-based CLUVI representation by means of their video clip identification tag, encoded both in the TMX file and in the video clip (in its file name).

These two sets of subtitled clips are stored as FLV files (because of their compression rate and small file size) in the server file system, where they are named –with a unique file name– according to their film title, their subtitle language (English or Galician), and their sequential number. Thus, whenever users search the Veiga they get both the bilingual text pair and the clips where this text/subtitle appears. On the other hand, the bitextual TMX files are stored with a file name according to their film title, and include the tags of both the in-cue and out-cue time of each subtitle and their sequential number. This information is encoded in the Veiga Corpus with a second element added to the TMX-based CLUVI tagging: the <s> element, which contains three attributes –s for the sequential number, d for the in-cue time, and a for out-cue time– for each tagged subtitle. To illustrate this, Listing 5 shows the code included in the TMX file named peixe.tmx (from the film entitled *Shooting fish*, by Stefan Schwartz) that would correspond to the video clips stored in the file system as peixe_en-848.flv, peixe_en-849.flv, peixe_gl-808.flv; and peixe_en-850.flv, and peixe_gl-809.flv.

Listing 5 Example of textual/audiovisual interface tagging in the Veiga Corpus.

```
<tu>
  <tuv xml:lang="en"><seg><s n="848" d="01:07:32,351" a="01:07:34,342"/>We
       play our cards right,<l/>we could end up with... <s n="849" d="01
      :07:34,431" a="01:07:36,023"/>two million pounds of tobacco<l/>to
      spend it for us.</seg></tuv>
  <tuv xml:lang="gl"><seg><s n="808" d="01:07:33,271" a="01:07:36,946"/>
      Podemos gastar 2 millóns en tabaco<l/>e pósters de Pamela Anderson.</
      seg></tuv>
</tu>
<tu>
  <tuv xml:lang="en"><seg><s n="850" d="01:07:36,511" a="01:07:39,025"/>I
      meant to get someone<l/>to spend it for us.</seg></tuv>
  <tuv xml:lang="gl"><seg><s n="809" d="01:07:37,071" a="01:07:39,539"/>
      Buscaremos alguén<l/>que o gaste por nós.</seg></tuv>
</tu>
```

4 Results

Since 2003, the SLI at the University of Vigo offers the possibility of searching and browsing the CLUVI parallel corpora online `http://sli.uvigo.es/CLUVI/`. The parallel corpora managed by the web application are stored in the XML CLUVI specification, whereas the searching and browsing tool designed in PHP was specifically created to carry out bilingual searches in tagged texts that are conformant to this specification. This search application allows for very complex searches of isolated words or sequences of words, and shows the bilingual equivalences of the terms in context, as they appear in real and referenced translations. Due to copyright issues, it returns a maximum of 1,500 hits only. Users can search terms in either language of the corpus, although it is also possible to carry out true bilingual searches, that is, to simultaneously search one term in each of the languages present in the parallel corpus. Search results are displayed in a parallel fashion as a list of translation units. In addition, the multimedia version of the Veiga Corpus has an improved browsing functionality that enables users to stream the subtitled video clips (stored as FLV files) via the open source video player Flowplayer[1], which allows embedding FLV video files into the results page. This multimedia-aware interface is already available for public consultation at`http://sli.uvigo.es/CLUVI/vmm.html`.

The coverage and size of the multimedia Veiga are shown in Table 1, where Words_EN means 'number of words in English', TUs stands for 'number of translation units' (roughly, English-Galician equivalent sentences), and Sub_EN means 'number of subtitles in English' (i.e. number of video clips subtitled in English).

4.1 Drawbacks

As [3] pointed out, we need to access texts in an in vivo form that provides access to audio and video tracks and maintains their relationship intact, because a major part of the way in which a film text makes its meaning is precisely through the synchronization between visual

[1] `http://flowplayer.org`

Table 1 Coverage and size of the Multimedia Veiga Corpus.

Film title	TUs	Words_EN	Words_GL	Sub_EN	Sub_GL
Afterglow	1300	6839	5811	911	1051
Babel	1208	6158	3925	941	777
Blood and Wine	1027	4353	3266	887	887
Bride of Chucky	1016	4750	4130	858	858
City of Industry	782	3391	3023	464	631
Earthlings	621	6822	6692	861	1097
Faces	2230	11631	7403	1566	1536
Fury	1481	9498	7268	1207	1189
If...	1093	5341	4464	869	879
Napoleon	6815	6075	1470	1212	933
Paris-Texas	1486	8549	5605	1138	1182
Punishment Park	1684	10140	8035	1494	1473
Shooting Fish	1636	9176	6699	1191	1146
	22379	92723	67791	13599	13639

and audio resources. However, before pinning any hope on the potential strengths of the Veiga corpus of subtitles, we had better begin by acknowledging its most visible weaknesses.

The first drawback is the small size of the corpus. To our credit, we must say that only two people are currently working on the project, and that a further extension is envisaged to also include TV broadcast films and other languages in a near future. A larger corpus would no doubt provide evidence of a wider range of phenomena, which may positively impact the reliability of any subsequent research based on the Veiga data. However, size is not necessarily a guarantee of representativeness. Moreover, in some circumstances, small, field-specific corpora may be equally useful for the investigation of particular phenomena.

The second limitation is the heterogeneous origin and authorship of the translated subtitles, which may accordingly call for different approaches to data observation and foreseeably arise the question of translation (and corpus) quality. As previously mentioned, the Galician subtitles were produced for DVD, cinema and Internet distribution. Specifically, seven of them are DVD-catered subtitles, that is, they are likely to be made by professional translators and to have undergone a quality control check. Fourteen of them were produced for the cinema. Notably, they were screened at various film series organized by a Galician film association. In this case, the subtitlers are mostly volunteers (non-paid translators), and they would lie halfway between the previous (professional) and the next (amateur) kind of translators. The other three sets of subtitles are instances of a new genre of subtitling in Spain (and other countries) that is properly known by the name of amateur subtitling and described by [5] as a practice 'undertaken by non-professionals and governed by dramatically different constraints than professional subtitling'. Often, the end result 'is conditioned by how much the subtitle producer has heard and understood from the original language', which is 'likely to result in a multitude of mistakes and misinterpretations'. Nonetheless, quality was not a criterion that we took into consideration when compiling our corpus.

And a third limitation is the above-mentioned processing and editing tasks involved in the process of creating a multimedia parallel corpus, which are still, and in spite of the technological advances, very time consuming. Consequently, no matter what purpose corpus users are driven by when searching the Veiga, they must keep these limitations in mind at all times.

4.2 Applications

Notwithstanding the aforementioned, our multimedia corpus may still serve a number of potential uses and purposes. First, it may be exploited as a reservoir of examples, offering researchers and scholars a database to analyse the different strategies and procedures used in both interlingual and intralingual subtitling and helping them substantiate their theoretical assumptions with practical evidence. From a pedagogical perspective, the Veiga features suggest that it could be used for different purposes in various learning settings, ranging from general language courses dealing with pronunciation, register, collocations, and other features of oral and written discourse, to specialised courses in audiovisual translation (AVT) with a focus on interlingual and intralingual subtitling ([17]). As put forward by [16], it is important that AVT teachers provide trainees with authentic material for contrastive analysis of both source (original) and target (translated) texts. Concerning language learning, the use of assorted 'real' texts, and particularly intralingual subtitles for L1 learning and interlingual subtitles for L2 learning, is likely to increase students' motivation and cultural awareness, although careful selection, adaptation and designing of teaching materials and activities coupled with adequate teacher guidance need to be in place. At the same time, the Veiga multimedia corpus may also prove a useful e-learning tool, since it would provide students with the possibility of exploring textual properties while listening to and watching film clips, which can be played and stopped at will ([2]), thus promoting autonomous learning. Finally, professional practitioners could also benefit from the possibility to access a collection of ready-made subtitles, where they can look at how other colleagues solved particular subtitling challenges.

As we have just discussed, the limited size of the corpus and the hybrid nature of the translated subtitles do not allow for generalizations about the practice of intralingual and interlingual subtitling. In fact, further distinctions could be made based on the particular genre of the audiovisual texts (featured films, documentaries, children's films...) and the product distribution medium. Nevertheless, corpus users should keep in mind that our core aim is to provide a tool that may serve not only researchers, but also practitioners and teachers to illustrate particular aspects of subtitling, and no regard is given to issues of corpus quality and representation.

On one hand, technical issues such as subtitles' display on screen (number of lines, alignment, position, colour, dialogue markers) and duration (in and out times, delay, shot change, synchronization) can be easily looked at in the Veiga corpus. The subtitling practice is rather heterogeneous and it can vary substantially from one audiovisual program, company and country to another ([6]). And although some efforts have been made to come up with a set of conventions or harmonized guidelines, subtitling tradition seems to determine what current practice is in each particular language/culture.

On the other hand, both interlingual and intralingual subtitles are condensed versions of the original audiovisual text. Subtitling usually involves the selection of linguistic material, forcing subtitlers to make decisions on what is important and what is seemingly superfluous or even redundant. Redundancy indeed is a very important concept in subtitling, because the information not given by the subtitles may be supplied by other elements present in the audiovisual text: the image and/or the sound ([4]). Reduction, however, is often achieved through the omission of information or by sacrificing interpersonal meaning ([12]). The Veiga multimedia corpus of subtitles not only places subtitles and the original audiovisual text in juxtaposition with one another, but also brings the English intralingual subtitles face to face with the Galician translated subtitles, allowing users to explore phenomena such as cohesion and condensation, which are deeply rooted in the semiotics of subtitling.

5 Conclusions

We have presented the Veiga multimedia corpus of English-Galician subtitles, an ongoing project that aims at echoing the general idea put forth by some authors, who claim to transcend the traditional only-text approach to corpus design and call for the need to build multimedia corpora that better reflects the polisemiotic aspects of film discourse and subtitling. After describing its content, we have raised some issues regarding corpus data and design, particularly the 'in-betweenness' of the data sets, which are not exactly translations of each other. In section 3 we briefly explain the methodology used to build the corpus, which partly mirrors that of all the other CLUVI's subcorpora, with the main difference being that the Veiga undergoes further processes in order to account for its audiovisual nature and to better cater to the principles of usefulness and usability. Then we sketch the main features of the corpus query system, which offers the option to stream the videoclips containing the subtitles. This is followed by an account of some obvious limitations of the corpus, such as size and technology constraints, that we hope to resolve in the near future. And finally, we have pointed at various areas in subtitling practice, research and education where the Veiga multimedia corpus could be of most value.

Right now, we are conducting some research concerning the potential applications of the Veiga multimedia in various pedagogical settings. As we move forward, we expect to improve the corpus in general and to illustrate its usage in other scenarios as well.

References

1 Mona Baker. Corpora in translation studies: An overview and some suggestions for future research. *Target*, 7(2):223–243, 1995.

2 Anthony Baldry. The role of multimodal concordancers in multimodal corpus linguistics. In Terry D. Royce and Wendy Bowcher, editors, *New Directions in the Analysis of Multimodal Discourse*, pages 173–193, Mahwah, NJ, 2006. Lawrence Erlbaum Associates.

3 Anthony Baldry and Christopher Taylor. Multimodal concordancing and subtitles with MCA. In Alan Partington, John Morley, and Louann Haarman, editors, *Corpora and Discourse*, pages 57–70, Bern, 2004. Peter Lang.

4 Francesca Bartrina. Teaching subtitling in a virtual environment. In Jorge Díaz Cintas and Gunilla Anderman, editors, *Audiovisual translation: Language Transfer on Screen*, pages 229–239, London, 2009. Palgrave Macmillan.

5 Jorge Díaz Cintas and Gunilla Anderman, editors. *Audiovisual translation: Language Transfer on Screen*. Palgrave Macmillan, London, 2009.

6 Jorge Díaz Cintas and Aline Remael. *Audiovisual Translation: Subtitling*. St. Jerome, Manchester, 2007.

7 Xavier Gómez Guinovart. A hybrid corpus-based approach to bilingual terminology extraction. In Isabel Moskowich-Spiegel Fandiño and Begoña Crespo, editors, *Encoding the Past, Decoding The Future: Corpora in the 21st Century*, pages 147–175, Newcastle upon Tyne, 2012. Cambridge Scholar Publishing.

8 Xavier Gómez Guinovart, Eva Díaz Rodríguez, and Alberto Álvarez Lugrís. Aplicacións da lexicografía bilingüe baseada en córpora na elaboración do Dicionario CLUVI inglés-galego. *Viceversa: Revista Galega de Traducción*, 14:71–87, 2008.

9 Xavier Gómez Guinovart and Elena Sacau Fontenla. Parallel corpora for the Galician language: building and processing of the CLUVI (Linguistic Corpus of the University of Vigo). In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pages 1179–1182, 2004.

10 Xavier Gómez Guinovart and Alberto Simões. Parallel corpus-based bilingual terminology extraction. In *Proceedings of the 8th International Conference on Terminology and Artificial Intelligence*, Toulouse, 2009. Université Paul Sabatier.

11 Xavier Gómez Guinovart and Alberto Simões. Translation dictionaries triangulation. In Carmen García Mateo, Francisco Campillo Díaz, and Francisco Méndez Pazó, editors, *Proceedings of FALA2010: VI Jornadas en Tecnología del Habla - II Iberian SLTech Workshop*, pages 171–174, Vigo, 2010. Universidade de Vigo.

12 Josélia Neves. Interlingual subtitling for the deaf (and hard-of-hearing). In Jorge Díaz Cintas and Gunilla Anderman, editors, *Audiovisual translation: Language Transfer on Screen*, pages 151–169, London, 2009. Palgrave Macmillan.

13 Yves Savourel and Arle Lommel. TMX 1.4b Specification. Technical report. Localisation Industry Standards Association. <`http://www.gala-global.org/oscarStandards/tmx/tmx14b.html`>, 2005.

14 Alberto Simões and Xavier Gómez Guinovart. Terminology extraction from English-Portuguese and English-Galician parallel corpora based on probabilistic translation dictionaries and bilingual syntactic patterns. In António Teixeira, Miguel Sales Dias, and Daniela Braga, editors, *Proceedings of the Iberian SLTech 2009 - I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages*, pages 13–16, Porto Salvo, 2009. Designeed.

15 Alberto Simões, Xavier Gómez Guinovart, and José João Almeida. Distributed translation memories implementation using WebServices. *Procesamiento del Lenguaje Natural*, 33:89–94, 2004.

16 Cristina Valentini. A multimedia database for the training of audiovisual translators. *The Journal of Specialized Translation*, 6:68–84, 2006.

17 Federico Zanettin, Silvia Bernardini, and Dominic Stewart, editors. *Corpora in Translator Education*. St. Jerome, Manchester, 2003.