

Using a Multimedia Parallel Corpus to Investigate English-Galician Subtitling

Patricia Sotelo Dios

Computational Linguistics Group
University of Vigo
Spain
psotelod@uvigo.es

Abstract

This paper presents an ongoing research project that involves the compilation and exploitation of the multimedia corpus of subtitled films *Veiga* as a method to investigate the practice of English intralingual subtitling and English-Galician interlingual subtitling. Our project draws on recent work in corpus-based translation studies and its applications in the field of audiovisual translation and, more specifically, on the use and development of multimedia corpora to empirically investigate film discourse, translation and subtitling. The *Veiga* corpus has been developed under the broader framework of the *CLUVI* Parallel Corpus, although it transcends the only-text approach that characterizes the *CLUVI* various corpora, enabling users to access the corpus content, that is, the English and the Galician subtitles of twenty four English speaking audiovisual products, in their natural, multi-semiotic form. The paper discusses issues of corpus design, multimedia text processing (alignment and annotation) and data retrieval alongside questions about the potential uses of our multimedia parallel corpus in subtitling and translation practice, research and education.

Introduction

Is it not a contradiction to set up a database or a corpus of film dialogues and their subtitles, with no pictures, and still pretend to study screen translation? (Gambier, 2006).

Quite a number of research projects concerning multimodal text analysis and the creation of multimodal corpora have been carried out in the last fifteen years. At the same time, both research on audiovisual translation and corpus-based translation studies have benefited from a rapid and prolific development. However, most of the studies on interlingual subtitling deal mainly with only linguistic and cultural matters, i.e. how to translate or convey cultural references, humour, dialects, register, taboo language, etc., neglecting the multisemiotic nature of the audiovisual text. As Gambier (2008) points out, certain concepts in translation studies should be revised, extended and rethought when they are applied to audiovisual translation, i.e. the concept of text, translation units, corpus design and content analysis, and a new methodology (and technology) to deal with multimodality is called upon. Because, if we consider films to be multimodal articulations of multiple and integrated discourses, as defined by Kress and van Leeuwen (2001), and we acknowledge that subtitles are not created to be read in isolation, the answer to Gambier's question is an affirmative one: it certainly would be a contradiction to set up a "text-only" corpus of subtitles if we aim at analyzing subtitles as transcriptions of film or TV dialogues, presented simultaneously on the screen, as defined by Baker & Saldanha (2009), and subtitling features such as time and space constraints.

Inspired by the new avenues for research offered by the exponential increase of data available in digital format and helped by the new methods, web and software tools, and other advances in multimedia corpus design, corpus developers and researchers can now tackle questions and attempt to find answers that would have been unthinkable using more traditional methods. A larger number of multimedia corpora of different genres and more theoretical and empirical studies are still needed in order to overcome current limitations in multimedia corpus building and exploitation.

In the first section we present the data that we have collected so far, and we describe the compilation, alignment and annotation procedures that have been followed to build the corpus, and its query system. The second section arises some discussion about the potential use of such a corpus as a tool for researchers, teachers and practitioners. And in the final section we draw some conclusions and highlight some challenges and future work.

The *Veiga* Corpus

The *Veiga* corpus is a small but ever-growing English-Galician corpus consisting of 24 American, British, and Australian films subtitled in both English (intralingual subtitling¹) and Galician (interlingual subtitling²) for DVD, cinema and Internet distribution. All in all, the corpus consists of approximately 300,000 words, but an extension is envisaged to include TV broadcast films and also other languages.

Developed under the broader framework of the *CLUVI* Parallel Corpus³, which is home to various parallel corpora of different language combinations and domains (law, science, technology...), 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 a reasonable way.

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' as these concepts have been traditionally defined in the literature. Baker (1995), for instance, uses the term 'parallel corpora' to refer to

¹ Intralingual subtitles are written in the same language as the original spoken dialogue. They are most often used by hard-of-hearing viewers and language learners.

² Interlingual subtitling is a type of language transfer in which the translation, that is the subtitles, do not replace the original source text, but rather, both are present in synchrony in the subtitled version (Georgakopoulou, 2009).

³ The *CLUVI* parallel corpus is available for public consultation at http://sli.uvigo.es/CLUVI/index_en.html.

"original, source-language texts in language A and their translated version in language B". In contrast, a comparable corpus (Sammut & Webb, 2010) can be defined as a document collection composed of two or more disjoint subsets, each written in a different language, such that documents in each subset are on a same topic as the documents in the others. The prototypical example of a comparable corpus is a collection of newspaper articles written in different languages and reporting about the same events: while they will not be, strictly speaking, the translation of one another, they will share most of the semantic content.

Thus, the *Veiga* corpus inhabits a certain no-man's 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, the 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. However, a parallel relation is very likely to come into play between the two sets of subtitles as well, inasmuch as they both are "sub-products" of the same original text. In sum, a double unidirectional parallel may be established between the original and the subtitles, and a bidirectional correlation is also expected to exist between the subtitles themselves, as shown in Figure 1.

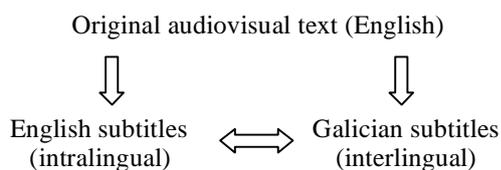


Figure 1: The subtitling triangle

Obviously, an only-text corpus of subtitles would not allow for any kind of parallel observance in terms of original 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 subtitling, e.g. phenomena related to the semiotics of interlingual and intralingual subtitling.

Corpus Design

As mentioned before, the *Veiga* corpus is hosted at the *CLUVI* (The Linguistic Corpus of the University of Vigo), which is being developed by the SLI (Computational Linguistics Group). The *CLUVI* functions as a repository of parallel subcorpora of different sizes and thematic fields, all of which undergo identical compiling and processing routines. This macrostructure implies that "corpus-makers" must follow the same alignment and

annotation procedures, which is an advantage for the working team and also for corpus users, who can access many corpora from one single web page with an identical display and user interface.

Nonetheless, the *Veiga* multimedia requires further processing: besides annotating events such as omissions, additions and reordering (less common in subtitling, on account of the need for synchrony) of translation units⁴, 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, including time and space constraints, segmentation, and condensation, among other particularities. In addition to this, the *Veiga* multimedia version⁵ incorporates a bonus feature: it enables users to stream the video clips where the bilingual pair appears, 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, there is a link to the corresponding video clips subtitled in each of the two languages (English and Galician).

Needless to say, the processes involved in the creation of a multimedia corpus are rather complex and very time consuming. However, the rapid advances in digital technologies and the Internet open up new possibilities to deal with multimedia and multimodal data. Consequently, we hope to witness an increase in the number of multimedia corpora in the near future. In what follows, we briefly describe the steps involved in designing, processing and browsing the *Veiga* multimedia.

Alignment, Annotation and Video-editing

Alignment is defined by Bowker (2002) as "the process of comparing a source text and its translation, matching the corresponding segments, and binding them together as translation units in a Translation Memory". The rationale behind parallel corpus alignment may include, but is not limited to, the intended usage of the corpus and the type of texts used as a database, whose segmentation level (word, sentence, paragraph...) shall vary accordingly.

As Guinovart & Sacau (2004) point out, the 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. Frequently one sentence of the source text corresponds with one sentence of the translation (1:1). Nevertheless, there are cases 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). This lack of correspondence between source and translated text (1:0, 0:1) is described as either omission or addition (always with respect to the source text) and tagged using an adapted version of TMX 1.4 content elements <hi> and <ph>. Regardless of the software tool used for aligning

⁴ For a thorough definition of translation unit (TU), see Baker & Saldanha (2009).

⁵ The *Veiga* multimedia corpus of subtitles is available for public consultation at http://sli.uvigo.es/CLUVI/vmm_en.htm.

Please note that the *Veiga* multimedia is still under construction, and only 10 of the 24 films are available in multimedia format at this writing.

text, all of the *CLUVI* subcorpora comply with this specification.

Film name (TU)	English subtitle	Galician subtitle
PAR (32)	They said they found Travis.	Atoparon a Travis.
PAR (33)	[[hi type='supr']]Oh, no. [[/hi]]	[[---]]
PAR (34)	What are you going to do?	-E que vas facer?

Table 1: Example of omission

Film name (TU)	English subtitle	Galician subtitle
BAB (1201)	In other news...	Noutras noticias...
BAB (1202)	[[---]]	[[hi type='incl']]Aos meus fillos, María Eladia e Eliseo... [[/hi]]
BAB (1203)	[[---]]	[[hi type='incl']]...as luces máis brillantes na noite máis escura. [[/hi]]

Table 2: Examples of insertion

Furthermore, due to the particular nature of our corpus, encoding of certain subtitling specificities such as subtitles' segmentation and duration was also included. As illustrated in Figure 2, the symbol ¶ indicates a line break within a subtitle, and when rolling over the film icon, information on the subtitle number and the in-cue and out-cue time will appear.

PEI (1204)	 We play our cards right , ¶ we could end up with...  two million pounds of tobacco ¶ to spend it for us. <small>n=848_d=01:07:32,351_a=01:07:34,342</small>
PEI (1205)	 I meant to get someone ¶ to spend it for us.

Figure 2: Subtitles' time code and line break annotation

Besides parallel alignment of translation units and annotation, which affect mainly the textual dimension of the corpus, some video-editing tasks are in force when compiling and processing audiovisual material. Considering that a) this is a corpus of subtitles, and b) one of the main constraints (and an object of study) of subtitling is time, a logical second step was to subject the corpus to a further segmentation and alignment stage, using subtitles as segmentation units. That is, all of the *Veiga* multimedia texts (English-language films) have been cut into video clips, each one corresponding to a subtitle. We still have two subsets of subtitles, the English and the Galician, each made up of as many videos as it has subtitles. Moreover, given that a high number of subtitles are not long enough (one, two seconds) to be played and watched properly, 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. Once we get two sets of subtitled clips for each audiovisual (AV) text, we link them to their corresponding text in the corpus (film

icons represent videos in Figure 2) so whenever users search the *Veiga* they get both the bilingual text pair and the clips where this text/subtitle appears. The major hurdle now is finding a freeware video-editing tool that offers customary, automatic batch splitting features.

In a few words, the *Veiga* multimedia corpus goes through two different segmentation processes: one affecting only the textual dimension, i.e. the subtitles, the other affecting the subtitles plus the original audiovisual text they come with. In the first case, the segmentation unit is the orthographical sentence (English sentence ↔ Galician sentence), while in the second case segmentation occurs at the subtitle level (English AV text + English subtitle ↔ English AV text + Galician subtitle). As for alignment, the relationship between English & Galician text units is of 1:1, 1:0 or 0:1 type, and these units do not necessarily match the subtitle segmentation units. Indeed, more often than not, we find that a single text unit is displayed in more than one clip/subtitle. As shown in Figure 3, a single English subtitle does not always map to a single Galician subtitle.

EAR (340)	Carbon monoxide poisoning, ¶ Strychnine, suffocation, ¶ breaking the neck, and anal electrocution ¶ are some of the more common methods used.	Envenenamento por ¶ monóxido de carbono, estricnina ¶ abafamento, ¶ quebrar o pescozo, ¶ e a electrocución ¶ anal ¶ son algúns dos métodos usados.
EAR (341)	Removed from his or her cage with a heavy ¶ neck-pole, the animal is walked past ¶ the rows of bodies of slaughtered foxes, ¶ sables, raccoons and wolves, among others.	¶ Sacado da súa gaiola ¶ cunha vara no pescozo, ¶ os animais camiñan pasando entre ¶ ringleiras de corpos de raposos, martas, ¶ mapaches, lobos e outros animais mortos.
EAR (342)	Death by anal electrocution is a crude ¶ process that requires a probe to be inserted ¶ in the rectum while the animal ¶ bites down on a metal conductor.	¶ A morte por electrocución anal ¶ é un proceso cruel, ¶ no cal se insere un electrodo no recto, ¶ mentres o animal traba ¶ nun condutor de metal.

Figure 3: Example of a search result. Non-1:1 subtitle correspondence

Data Retrieval

As previously mentioned, both the *CLUVI* Parallel Corpus and the *Veiga* multimedia subcorpus can be searched online via a PHP application designed by the SLI. One of the key features of this search tool is that it has a quite comprehensive query language. It allows for complex as well as parallel searches of isolated words or sequences of words, and shows the bilingual equivalences of the searched terms in context. Searches can be made from and to either of the languages (English or Galician) or from both languages, that is, users can simultaneously search for a term in each of the two languages (English and Galician). All the *CLUVI* bilingual corpora use the same search interface⁶: a simple box where users can enter their query in any or all of the languages. Due to copyright issues, it returns a maximum of 1.500 hits only.

1-CHU (195)	When you're 18, you can ¶ go to hell ¶ for all I care.	Canda teñas 18, ¶ por min ¶ como se vas para o inferno,		
2-CHU (258)	I've been a prisoner of my love ¶ for you ¶ for a very long time.	Levo moito tempo presa ¶ no amor que sinto ¶ por ti.		
3-CHU (436)	I'd do anything ¶ for you.	¶ - Faría calquera cousa ¶ por ti,		

Figure 4: Example of a search result

For example, we searched for bilingual pairs where the word 'for' in the English and the word 'por' in the Galician appear. The search engine returned a list of 131 units that matched our query. As we can see in Figure 4, results are

⁶ Note: The *Veiga* multimedia is currently migrating to a new web site.

displayed in a five-column table that contains the following information: film code and translation unit number, English unit, Galician unit, movie reel icon, and arrow icon. The arrow allows users to access the co-text. When clicked, a new window pops up showing this specific pair together with the previous and the next segments. And if we click on the reel icon, we are automatically directed to a page where we can play the corresponding video streams in the same parallel fashion. Again, both the previous and the next video/subtitle are provided.

Potential Uses and Limitations of the *Veiga* Multimedia

Baldry (2004) already stated it: 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 and audio resources. However, before pinning any hope on the potential strengths of the *Veiga* multimedia corpus of subtitles, we'd better begin by acknowledging its most visible weaknesses.

Limitations

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, as McEnery & Wilson (2001) pointed out, 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 we have said before, 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 some quality control. 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 subtitles 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 set of subtitles are instances of a new kind of subtitling in Spain (and other countries) that is properly known by the name of amateur subtitling and described by Díaz Cintas & Anderman (2009) 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.

Potential Uses

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 interlingual and intralingual subtitling and helping them substantiate their theoretical assumptions with practical evidence. From a pedagogical perspective, it could be used in different contexts, ranging from general language courses dealing with jargon and register to specialised courses on audiovisual translation (AVT) and subtitling. As Valentini (2006) suggests, it is important that AVT teachers provide trainees with authentic material for contrastive analysis of both source (original) and target (translated) texts. At the same time, it 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 (Baldry, 2006). 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 medium used for the distribution of the product. Nevertheless, it should be clear that our only aim is to provide a tool that may serve not only researchers, but also practitioners and teachers to illustrate particular aspects of subtitling.

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, delay⁷, shot changes, synchronization) can be easily looked at in the *Veiga* corpus. According to Díaz Cintas & Remael (2007), the subtitling practice is rather heterogeneous and it can vary substantially from one audiovisual program, company and country to another. 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.

⁷ De Linde & Kay (1999) distinguish between 'leading' and 'lagging'. If a subtitle precedes the speech the delay is referred to as lead time, if a subtitle follows speech it is referred to as lag time.

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 (Bartrina, 2009). Reduction, however, is often achieved through the omission of information or by sacrificing interpersonal meaning (Neves, 2009). 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.

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. 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. 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. Finally, we have pointed at various areas in subtitling practice, research and education where the *Veiga* multimedia corpus could be of most value.

Acknowledgements

This work was developed within the framework of the project *Desenvolvemento e explotación de recursos integrados da lingua galega*, which was funded by *Consellería de Innovación e Industria, Xunta de Galicia* under grant INCITE08PXIB302185PR. It was carried out at the Computational Linguistics Group, which is part of the Research Group on Galician Language Technology and Applications (TALG Group) of the University of Vigo.

References

Baker, M. & Saldanha, G. (Eds.) (2009). Routledge Encyclopedia of Translation Studies, Second Edition (pp. 244-245, 304-306). London & New York: Routledge.

— (1995). Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target*, 7 (2), 223-243.

Baldry, A. (2006). The Role of Multimodal Concordancers in Multimodal Corpus Linguistics. In T. Royce & W. Bowcher (Eds.), *New Directions in the Analysis of Multimodal Discourse* (pp. 173-193). Mahwah, NJ: Lawrence Erlbaum Associates.

Baldry A. & Taylor, C. (2004). Multimodal concordancing and subtitles with MCA. In A.

Partington, J. Morley, & L. Haarman (Eds.), *Corpora and Discourse* (pp. 57-70). Bern: Peter Lang.

Bartrina, F. (2009). Teaching Subtitling in a Virtual Environment. In J. Díaz Cintas & G. Anderman (Eds.), *Audiovisual translation: Language Transfer on Screen*. London: Palgrave Macmillan.

Bowker, L. (2002). Computer-Aided translation Technology: A Practical Introduction (p. 109). Ottawa: University of Ottawa Press.

De Linde, Z. & Kay N. (1999). *The Semiotics of Subtitling*. Manchester: St. Jerome.

Díaz Cintas, J. & Anderman, G. (Eds.) (2009). *Audiovisual translation: Language Transfer on Screen* (pp. 1-17). London: Palgrave Macmillan.

Díaz Cintas, J. & Remael, A. (2007). *Audiovisual Translation: Subtitling*. Manchester: St. Jerome.

Gambier, Y. (2008). Recent developments and challenges in audiovisual translation research. In D. Chiaro, C. Heiss, & C. Bucaria (Eds.), *Between Text and Image*. Amsterdam/Philadelphia: John Benjamins Publishing.

— (2006). *Multimodality and Audiovisual Translation*. In *MuTra 2006: Audiovisual Translation Scenarios: Conference Proceedings*. Copenhagen. Available online at http://www.euroconferences.info/proceedings/2006_Proceedings/2006_Gambier_Yves.pdf.

Georgakopoulou, P. (2009). Subtitling for the DVD Industry. In J. Díaz Cintas & G. Anderman (Eds.), *Audiovisual translation: Language Transfer on Screen*. London: Palgrave Macmillan.

Gómez Guinovart, X. & Sacau Fontenla, E. (2004). Parallel corpora for the Galician language: building and processing of the CLUVI (Linguistic Corpus of the University of Vigo). In T. Lino et al. (Eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)* (pp. 1179-1182). Lisbon. Available online at <http://webs.uvigo.es/sli/arquivos/lrec2004.pdf>.

Kress, G. & van Leeuwen, T. (2001). *Multimodal Discourse: The Modes and Media of Contemporary Communication*. New York: Oxford University Press.

McEnery, T. & Wilson, A. (2001). *Corpus Linguistics*, Second Edition. Edinburgh: Edinburgh University Press.

Neves, J. (2009). Interlingual Subtitling for the Deaf (and Hard-of-Hearing). In J. Díaz Cintas & G. Anderman (Eds.), *Audiovisual translation: Language Transfer on Screen*. London: Palgrave Macmillan.

Sammut, C. & Webb, G.I. (Eds.) (2010). *Encyclopedia of Machine Learning*. New York: Springer.

Valentini, C. (2006). A Multimedia Database for the Training of Audiovisual Translators. *The Journal of Specialized Translation*, 6, 68-84. Available online at http://www.jostrans.org/issue06/art_valentini.php.