Multilingual Videos for MOOCs and OER

Juan Daniel Valor Miró[1], Pau Baquero-Arnal[1], Jorge Civera[1*], Carlos Turró[2] and Alfons Juan[1]

[1]MLLP, DSIC, Universitat Politècnica de València, València, Spain // [2]Media Services Universitat Politècnica de València, València, Spain // jvalor@dsic.upv.es // jcivera@dsic.upv.es // ajuan@dsic.upv.es // pabaar@inf.upv.es // turro@cc.upv.es

[*]Corresponding author

(Submitted September 30, 2016; Revised November 23, 2016; Accepted January 1, 2017)

ABSTRACT

Massive Open Online Courses (MOOCs) and Open Educational Resources (OER) are rapidly growing, but are not usually offered in multiple languages due to the lack of cost-effective solutions to translate the different objects comprising them and particularly videos. However, current state-of-the-art automatic speech recognition (ASR) and machine translation (MT) techniques have reached a level of maturity which opens the possibility of producing multilingual video subtitles of publishable quality at low cost. This work summarizes authors' experience in exploring this possibility in two real-life case studies: a MOOC platform and a large video lecture repository. Apart from describing the systems, tools and integration components employed for such purpose, a comprehensive evaluation of the results achieved is provided in terms of quality and efficiency. More precisely, it is shown that draft multilingual subtitles produced by domain-adapted ASR/MT systems reach a level of accuracy that make them worth post-editing, instead of generating them ex novo, saving approximately 25%–75% of the time. Finally, the results reported on user multilingual data consumption reflect that multilingual subtitles have had a very positive impact in our case studies boosting student enrolment, in the case of the MOOC platform, by 70% relative.

Keywords

Video lecture repositories, MOOCs, Speech recognition, Machine translation, Multilingual

Introduction

Massive Open Online Courses (MOOCs) are rapidly growing since 2011, with more than 35 million students and 4000 courses offered at the beginning of 2016, roughly doubling the figures of the previous year (Shad 2015). Although US-based providers like edX and Coursera are now targeting international students, most courses are only delivered in English (76%), Spanish (8%), French (5%) or Chinese (3%) (see class-central.com/languages). For MOOCs to reach a worldwide audience, they should be provided in multilingual form. And this also holds true for Open Educational Resources (OER) in general. Although MOOCs and OER comprise objects of different kinds, in this work we focus our attention on producing multilingual video lectures; that is, on adding subtitles in their source (spoken) language(s) and then translate them into different target languages. Apart from its application to MOOCs and OER, multilinguality is of great interest in all contexts where educational videos are used. This includes online education in general (Kay, 2012), flipped teaching (Bishop & Verleger, 2013), and in-class recording services (Ketterl et al., 2010).

A direct approach to obtain source video subtitles is to generate automatic transcriptions by using *Automatic Speech Recognition (ASR)* technology. Indeed, the application of ASR technology to lecture recordings is by no means new. A detailed account of significant efforts in this domain up to 2010 can be found in (de-Pablos et al., 2011). More recent research efforts on ASR applied to educational videos can be found in the European projects transLectures (*Transcription and Translation of Video Lectures*) and EMMA (*European Multiple MOOC Aggregator*) (see platform.europeanmoocs.eu). Broadly speaking, from the results of these efforts we may conclude that ASR technology has reached a level of maturity that allows us to generate low-cost, automatic source subtitles of (nearly) publishable quality in most cases. It is worth noting, however, that such quality is only achievable by developing state-of-the-art ASR systems adapted to the particular task (media repository) at hand. In comparison with mainstream providers (e.g., YouTube), adapted systems achieve relative accuracy improvements of about 40%. In any case, even if automatic source subtitles are of moderate quality, they are often very useful for different purposes such as improving accessibility for hearing-impaired and foreign students (de-Pablos et al., 2011; Ranchal et al., 2013), video-clip search based on keywords (Repp et al., 2008) and discovery of content-related videos in a repository (Glass et al., 2007).

Analogously to the case of source subtitles, a direct approach to obtain target video subtitles is to generate automatic translations by using *Machine Translation (MT)* technology. This approach has been also explored with good results in transLectures and EMMA, and more recently in TraMOOC (Kordoni et al., 2016). A clear

conclusion from these results is that the translation quality of adapted MT systems is often accurate enough for post-editing; that is, it is often the case that the automatic translation is not far from the correct translation, and thus it is more time-efficient to review it than to produce the entire translation manually. In addition to this, as in ASR, system adaptation has been shown to be a key factor in maximizing output quality: in comparison with mainstream providers (e.g., Google Translate), adapted MT systems increase translation quality by about 20% relative. MT is normally applied to clean, post-edited automatic transcriptions and, as indicated above, automatic translations are also post-edited to end up with target subtitles of publishable quality. Regarding this, it is worth noting that many approaches have been considered to increase user productivity when reviewing subtitles, but post-editing is still the most popular (Plitt & Masselot, 2010; Specia, 2011; O'Brien & Simard, 2014; Valor-Miró et al., 2015).

The above discussion does not mean that the task of producing multilingual videos for MOOCs and OER simply comes down to developing advanced ASR/MT systems for lecture recordings. Obviously, it requires expertise, resources and tools from ASR/MT, but also additional components and experience for their proper integration into real-life educational environments. In this respect, this article summarizes a large part of the experience gained on this task by the Universitat Politècnica de València (UPV)'s *Machine Learning and Language Processing (MLLP)* group during transLectures and EMMA. Our main goal is to provide a comprehensive evaluation of the results achieved in a real-life MOOC platform and a large video lecture repository. However, resources and tools are described in broad terms since our focus here is not on ASR/MT technical details. On the contrary, here we report detailed results in terms of quality and efficiency, as well as on the impact multilingual videos have had in our real-life case studies.

The article is organized as follows. After a review of our case studies, the systems, tools and integration components required for multilingual video production are summarized. Then, detailed results on transcription and translation quality are provided, also including comparative results with mainstream providers. These results are followed by a thorough evaluation of transcription (translation) reviewing time for each language (language pair) considered separately, and also across all languages considered. Next, the impact these systems, tools and integration components have had in the case studies. Finally, the main conclusions drawn are summarized.

Case studies

This section introduces two case studies in which multilingual video subtitles are delivered: a MOOC platform and a large video lecture repository.

The EMMA platform

The European project EMMA (February 2014 – July 2016) involved 12 partners delivering more than 30 multilingual MOOCs on diverse topics.

Figure 1. Screenshot of a trilingual MOOC

Multilingualism is a distinctive feature of the EMMA platform as it provides built-in automated transcription and translation for all video and text contents. This includes transcription in 7 languages (English, Italian, Spanish, Dutch, French, Portuguese and Estonian) and automatic translation into English, Spanish and Italian. Automatic transcriptions and translations are reviewed by lecturers to reach publishable quality. Most courses have been offered in bilingual (original language plus English) or trilingual form (with Spanish, French or Italian as a third language). Figure 1 shows a unit of a trilingual MOOC in French then translated into English and Italian. A translation button allows to switch between languages.

The UPV media repository

The UPV media repository is a service for the creation, storage, management and dissemination of video lectures, called poliMedias. poliMedias provide a concise overview of a given topic and have an average duration of ten minutes (Turró et al., 2009). Figure 2 shows an example with subtitles in Spanish and English. Table 1 shows basic statistics on poliMedias by their most common languages. poliMedia subtitles can be reviewed anonymously, though editions must be approved by the lecturer before publication.

Table 1. Number of poliMedia hours of video per language

Language	Videos	Hours	Lecturers
Spanish	15013	2709	1572
English	1221	173	203
Catalan	434	52	80

Figure 2. A poliMedia with subtitles in Spanish and English

Systems, tools and integration components

Two main open-source tools have been used to develop ASR and MT systems: the *transLectures-UPV Toolkit (TLK)* (del-Agua et al., 2014) and the Moses toolkit (Koehn et al., 2007). These systems have been adapted by applying the techniques described in Martínez-Villaronga et al. (2013) and Axelrod et al. (2011). Then, these systems have been integrated into the case studies using the *transLectures-UPV Platform (TLP)* (see mllp.upv.es/tlp).

Figure 3 shows two examples of use of the multilingual TLP player/editor. The first example (top) is an editor for transcriptions. The video and its segmentation are displayed to the left, while transcriptions are shown to the right. The second example (bottom) is an editor for transcriptions and translations. It is analogous to the first example, but with translations also available to the right. Also, the TLP player/editor can be used to review text documents.

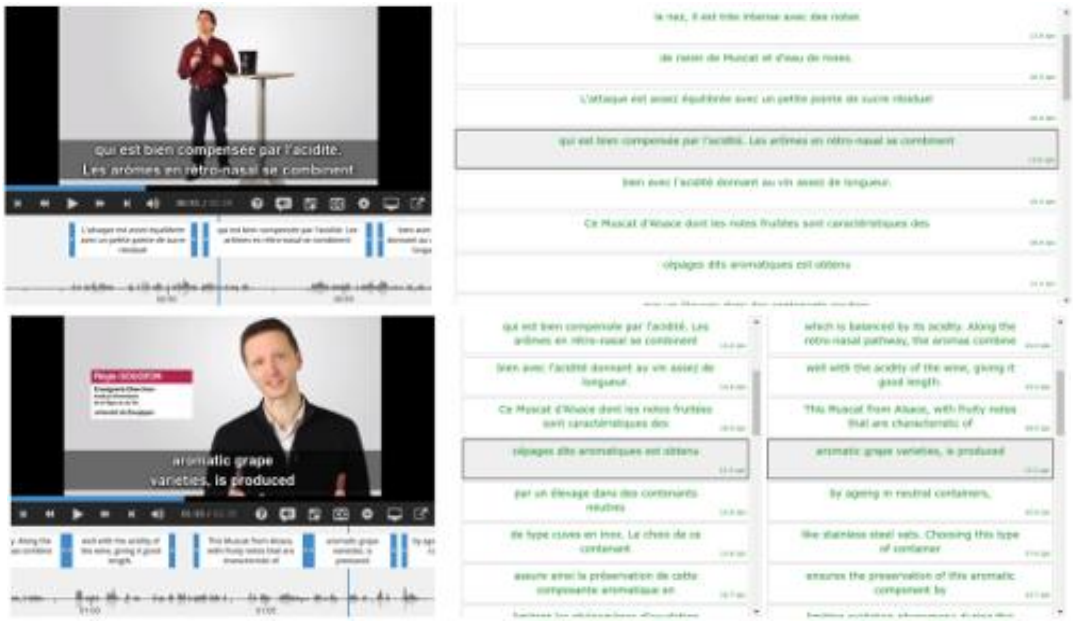

Figure 3. Multilingual editor for transcriptions (top) and translations (bottom)

The systems developed by the MLLP research group for the EMMA platform and the UPV media repository can be freely tried through the *Transcription and Translation Platform (TTP)* (see ttp.mllp.upv.es).

Transcription and translation quality

In this section, we assess the quality of automatic transcriptions and translations generated by the MLLP's ASR/MT systems for videos originally in 5 languages drawn from the UPV media repository and the EMMA platform. Additionally, a comparative evaluation of transcription and translation quality with mainstream providers of ASR/MT technology, i.e., YouTube and Google Translate, is also presented.

Transcription quality

Transcription quality was measured with the widely accepted *Word Error Rate (WER)* criterion (Hunt, 1990). Formally, the WER is the normalized minimum number of elementary word editing operations required to transform an automatic transcription into its corrected (reviewed) version. Three elementary word editing operations are considered: insertions, deletions and substitutions. Normalization is computed with respect to the number of words in the reviewed transcription, and often expressed as a percentage. For example, if a lecturer has to apply 30 elementary editing operations to an automatic transcription so as to obtain a reviewed version with a length of 200 words, then the WER will be 15%. In this regard, it must be noted that expecting to achieve error-free transcriptions is unrealistic, even if they are manually produced. On the contrary, it is more realistic to expect a WER of about 10% from commercial, manual transcription services (Hazen, 2006). From a practical point of view, automatic transcriptions of WER equal or less than 25% convey enough correct information to be useful (Munteanu et al., 2006), and professional stenographers prefer them to manually transcribing from scratch (Akita et al., 2009).

Table 2 shows the number of videos, duration (in hours) and WER (± standard deviation) for each transcribed language. Spanish- and English-language videos come from the UPV media repository, while Italian-, Dutch- and French-language videos were included in MOOCs delivered on the EMMA platform. There are a significant number of Spanish-language videos, since more than 90% of the videos in the UPV media repository are in Spanish.

The average duration of videos for all languages except for Dutch is less than 10 minutes. Dutch videos last more than 35 minutes on average and the format of the video presentation is different from that in the other languages. Dutch videos are interviews with usually two speakers sitting around a table, while in the other videos a single speaker stands in front of the camera.

Table 2. Videos, duration (hrs.) and WER (± std. dev.) per language

Language	Videos	Hours	WER
Spanish	207	24.7	18.4 ± 6.4
Italian	13	1.2	25.7 ± 6.4
English	25	3.5	21.9 ± 8.5
Dutch	11	6.9	29.4 ± 9.2
French	21	2.1	23.2 ± 8.3

From the results in Table 2, we can observe that the quality of Spanish transcriptions is the highest, followed by English and French, all three being below 25%. Italian is just above 25% of WER and Dutch has the highest WER figure, but still below 30% of WER. In the case of Dutch, we believe that the higher WER figure is explained by the presence of more than one speaker in the videos, which harms the acoustic adaptation to the speaker, not being so effective as in the rest of the videos in which a single speaker appears.

Translation quality

As with transcription, translation quality is often measured with an error criterion: the so-called *Translation Edit Rate* (TER) (Snover et al., 2006). This criterion is computed in the same way as the WER, which is, as a normalized percentage of the minimum number of elementary word editing operations required to transform an automatic output (translation) into its reviewed version. The only significant difference is that, apart from insertions, deletions and substitutions, shifts are also allowed. Also as with the WER, it must be noted that achieving error-free translations, either automatic or manual, is unrealistic. Additionally, in the case of MT it is generally accepted that source sentences can be manually translated in many different yet correct ways, and thus a correct translation for a certain reviewer might not be the preferred (correct) translation for another one. As the TER is computed from only one correct reference, it is considered a pessimistic criterion. From a practical point of view, automatic translations with TER figures below 50% are worth post-editing, instead of translating from scratch (Specia et al., 2009; Specia, 2011).

Table 3 shows the number of videos, duration (in hours) and TER (± standard deviation) for each translation pair. All videos were automatically translated and then reviewed. The Spanish-language videos are part of the UPV media repository and were reviewed by lecturers. The English→Spanish videos are from two EMMA MOOCs originally in Italian, then translated into English, and now for this work translated into Spanish. Analogously, the English→Italian videos are from two EMMA MOOCs originally in Spanish, then translated into English, and finally translated into Italian. In this evaluation set there are four MOOCs available in three languages (Italian, English and Spanish). Finally, the Dutch- and French-language videos are also from EMMA MOOCs translated into English.

Table 3. Videos, duration (hrs.) and TER (± std. dev.) per translation pair

Translation pair	Videos	Hours	TER
Spanish → English	101	10.8	33.2 ± 14.4
English → Spanish	29	2.5	27.0 ± 19.9
Italian → English	14	1.6	37.5 ± 8.2
English → Italian	121	6.5	33.8 ± 8.0
Dutch → English	5	3.5	30.7 ± 13.4
French → English	8	0.9	58.9 ± 5.2

From the results in Table 3, it is clear that, apart from the French→English pair, the translation quality is good enough to be worth post-editing (below 50% TER). The translation quality of the French→English pair was lower than expected. This phenomenon is due mainly to two reasons. First, the reviewers used a two-pass review process when generating the final translations that makes them differ significantly from those that would be obtained in a single pass, as was the case with the other translation pairs. Second, we believe that the MT system providing the automatic English translations from French did not properly adapt to the specific domain of the French courses.

Comparison with mainstream providers

One of the questions that arises is how the adapted systems deployed in this work compare to systems from mainstream providers and, in particular, to the state-of-the-art YouTube automatic captioning and Google

Translate systems. For this purpose, a different evaluation set was defined with videos from MOOCs offered in the EMMA platform. Table 4 shows, for each transcribed language, the number of videos included in this evaluation set, their duration, and the WER achieved by the MLLP's TTP and YouTube's automatic captioning.

Table 4. Videos, duration (hrs.), and TTP and YouTube WER per language

Language	Videos	Hours	TTP	YouTube
Spanish	23	3.5	14.8	22.5
Italian	3	4.0	17.1	31.6
English	9	0.4	39.2	65.9
Dutch	2	1.1	24.5	41.1
French	18	2.3	20.6	32.0

From the results in Table 4, we can conclude that YouTube's WER is higher than that of TTP's systems for all languages, and more precisely, the relative WER increase over TTP's is nearly 70% on average. The main reason behind these results is the fact that YouTube uses general-purpose ASR systems, while the ASR systems integrated into TTP are automatically adapted to the task as described in Section 3. The English ASR system obtained a surprisingly high WER compared with the one reported in Table 2 based on the same technology. An error analysis on the English videos studied in this work led to the conclusion that the accent of the only speaker in these videos was especially difficult to understand.

The evaluation set used to compare transcriptions was enlarged for the purpose of comparing translations. Table 5 shows, for each translation pair, the number of videos included in the translation evaluation set, their duration, and the TER obtained with the MLLP's TTP and Google Translate. These videos were previously transcribed in order to be translated. In the case of English into Spanish and French into English, the same English- and French-language videos transcribed in Table 4 were then translated.

Table 5. Videos, duration (hrs.), and TTP and Google TER per translation pair

Translation pair	Videos	Hours	TTP	Google
Spanish → English	250	13.9	33.9	44.3
English → Spanish	9	0.4	35.8	42.4
Italian → English	11	1.1	33.4	39.2
English → Italian	81	5.4	39.7	43.3
Dutch → English	2	1.2	42.5	45.0
French → English	18	2.3	52.8	52.6

A general conclusion that can be drawn from Table 5 is that Google Translate's MT systems produce a higher translation error than TTP's MT systems, except for French into English, where both systems show a similar performance. On average, the TER figures achieved by Google Translate are higher than those of TTP by 14% relative. Again, as opposed to the general-purpose MT systems provided by Google Translate, TTP systems are adapted to the domain of the video that is being translated, and thus more accurate results are obtained.

Reviewing time

The time required for reviewers (e.g., lecturers) to post-edit automatic video transcriptions and translations is measured in terms of *Real Time Factor (RTF)* (Valor-Miró et al., 2015). This measure is the video duration-normalized time required for the reviewer to post-edit the whole video transcription (or translation). For instance, if a video lasts 6 minutes and its review takes one hour (60 minutes), then the RTF will be 10.

In general, manual annotation of speech ranges from 10 RTF, in the case of orthographic transcription (Reidsma et al., 2005), to 50 RTF, in which a detailed 4-level speech annotation is performed (Barras et al., 2001). Expert transcriptionists can achieve as low an RTF as 6 (Williams et al., 2011), but this is not the usual profile for lecturers. In our previous work, (Valor-Miró et al., 2015), the RTF for manual (orthographic) transcription attained by lecturers was 10.1 ± 1.8, which matches the figures reported in (Reidsma et al., 2005). For this reason, we take 10 RTF as a reference review time for transcription.

Regarding the RTF for translation, in contrast to transcription, it is more difficult to establish a single reference RTF, except for the rule of thumb of 2500 words per day of work, since translation is a more complex task requiring a greater cognitive effort and involving different factors such as source and target languages, degree of expertise and experience of the translator, vocabulary specificity, software tools, etc. Having in mind this

limitation, specialist translators achieve fully-manual translating rates ranging from 400 to almost 1000 words per hour (Plitt & Masselot, 2010). Taking these figures into the UPV media repository in which speakers utter 150 words per minute on average, a specialist translator would be translating at 22.5 RTF in the worst case. In the transLectures project (Turró et al., 2016), seven hours of videos drawn from the UPV media repository were translated ex novo from Spanish into English by two professional translators achieving an average RTF of 34.1 ± 11.4 RTF. For the sake of comparison and taking into account the profile of the translators in this case (lecturers), hereinafter we consider the RTF of manual translation to be 30 RTF.

Transcription reviewing time

Table 6 shows, for each transcribed language, the average WER (copied from Table 2) and RTF (± std. dev.), and regression models to predict RTF as a function of WER. Three regression models were tried: linear, square root and logarithm. In the case of Spanish, detailed information is provided in Table 6 on the adjustment of these three regression models. Also, Figure 4 shows a scatter plot of RTF (y axis) versus WER (x axis) for each Spanish-language video (plotted point) and each adjusted regression model. For the rest of the transcribed languages, only the details on the adjustment of the logarithmic model are given in Table 6 for brevity.

Table 6. Average WER and RTF (± std. dev.), and regression models per language

Language	WER	RTF	Model	R^2	β	*Sig.*
Spanish	18.4	3.3 ± 1.2	WER	0.87	0.17	$< 10^{-15}$
			√WER	0.90	0.78	$< 10^{-15}$
			ln WER	0.91	1.17	$< 10^{-15}$
English	21.9	5.3 ± 1.7	ln WER	0.92	1.76	$< 10^{-14}$
Italian	25.7	3.9 ± 1.4	ln WER	0.90	1.20	$< 10^{-6}$
Dutch	29.4	5.8 ± 2.5	ln WER	0.85	1.75	$< 10^{-14}$
French	23.2	6.7 ± 0.8	ln WER	0.98	2.17	$< 10^{-15}$

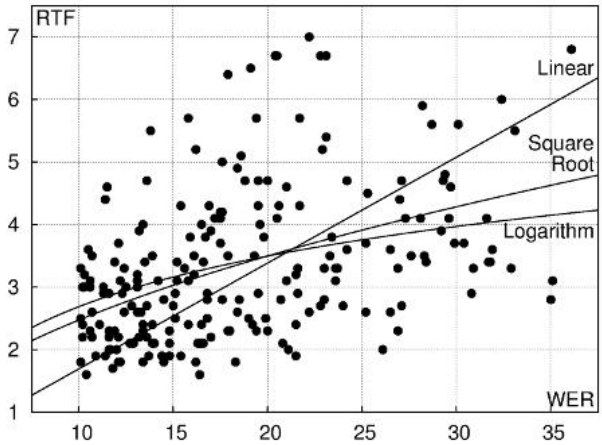

Figure 4. RTF vs. WER for Spanish-language videos and prediction models

A first important conclusion from the results on transcription reviewing time is that the availability of automatic transcriptions reduces between one third and two thirds the time devoted to generate video transcriptions. Generally speaking, we may say that the RTF is between 3 and 7 when starting from automatic transcriptions that are worth post-editing, as shown in Table 6. The second important conclusion is that the logarithmic regression model provides a good, statistically significant fit of the observed data, better indeed than the other two models considered. The logarithmic model explains better the fact that users tend to ignore automatic transcriptions when the corresponding WER is too high and prefer retranscribing from scratch to correcting a low-quality automatic transcription. For all languages, the adjustment is statistically significant (*Sig.* $< 10^{-4}$) and an important amount of the variability of the data is explained by the model ($R^2 \geq 0.85$).

On a per-language analysis, Dutch presents higher RTF figures than Spanish, Italian and English. We believe this is explained by the interview format of these videos. Finally, the RTF figure for French is not the one expected from the WER figure reported; indeed, this RTF figure is the highest in this transcription evaluation. The reason behind this RTF figure is the two-pass review process that lecturers carried out in this case. The second pass in the review process requires at least 1 additional RTF, which is the minimum amount of time required to watch the entire video again.

Translation review time

Table 7 shows, for each translation pair, the average TER (copied from Table 3) and RTF (± std. dev.), and regression models to predict RTF as a function of TER. Translation results are provided in Table 7 and Figure 5, in the same way as above for transcription.

Table 7. Average TER and RTF (± std. dev.), and regression models per translation pair

Translation pair	TER	RTF	Model	R^2	β	*sig*
Spanish → English	33.2	9.1 ± 4.9	TER	0.75	0.25	$< 10^{-15}$
			√TER	0.80	1.61	$< 10^{-15}$
			ln TER	0.80	2.71	$< 10^{-15}$
English → Spanish	27.0	7.8 ± 4.9	ln TER	0.82	2.67	$< 10^{-11}$
Italian → English	37.5	11.3 ± 4.2	ln TER	0.89	3.15	$< 10^{-7}$
English → Italian	33.8	9.6 ± 5.3	ln TER	0.77	2.76	$< 10^{-15}$
Dutch → English	30.7	9.5 ± 3.9	ln TER	0.91	2.89	$< 10^{-2}$
French → English	58.9	23.2 ± 8.0	ln TER	0.90	5.67	$< 10^{-4}$

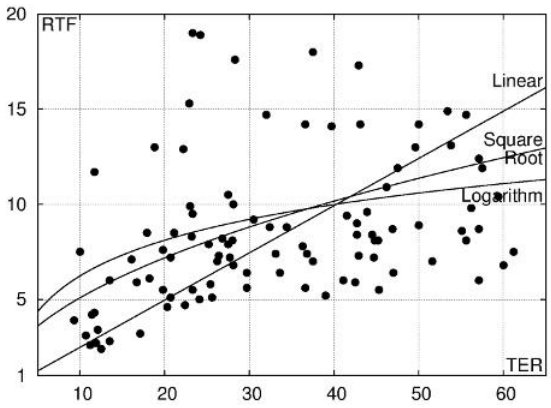

Figure 5. RTF vs. TER for Spanish into English videos; and prediction models

Similarly to transcription, the first important result is that, except for French→English, the review time is reduced to approximately one third when the quality of the automatic translations is worth post-editing, as shown in Table 7. The second result is that the logarithmic regression model is among the best explaining the observed data. Again, the logarithmic model fits better the cases with high values of TER, bounding the corresponding RTF, since reviewers ignore those automatic translations containing too many errors and prefer to generate the translation from scratch. The amount of the variability of the data explained by the model (R^2 values) is not as high as in the review of transcriptions, which is reflected in Figure 5 as a greater dispersion of the data points. The reason behind this behaviour is the higher complexity of the translation task (compared to transcription), which involves a significant cognitive load.

In a per-translation-pair analysis, the review of Spanish translations from English transcriptions is similar to the translation in the opposite translation pair, but the RTF figure is even lower for the latter. This fact correlates with the Italian into English and English into Italian translation pairs, since most of the reviewers involved are non-native English speakers, and it is easier for them to translate into their mother tongue. The figures for the Dutch into English translation review are very much in line with the previous translation pairs, considering that the quality of the automatic translations was among the best. Finally, the translation of French courses was surprisingly cumbersome, taking far more time than the other translation pairs. This phenomenon is due mainly to two reasons. First, as mentioned above, the MT system that generated the automatic English translations from French did not properly adapt to the domain of the courses; and second, reviewers employed a two-pass review process that was more costly than the conventional one-pass review process used in the rest of translation pairs.

Review time across languages

In the previous sections we have found that, for each language involved, a logarithmic regression model can be adjusted to accurately predict RTF from transcription WER; and we have reached a similar conclusion in translation (i.e., to predict RTF from TER) for each translation pair assessed. Therefore, it is worth asking whether a single logarithmic regression model could suffice to accurately predict RTF from WER (TER) across

all languages (translation pairs) under study. This is considered in Figure 6. The scatter plot at the top shows RTF versus WER, for all languages involved (plotted points), and a single logarithmic regression model fitted to data (videos) pooled across languages. The scatter plot at its bottom is similar, but for TER.

As for predicting RTF from transcription WER, the fitted logarithmic model shown at the top of Figure 6 (R^2 = 0.87, β = 1.34) is statistically significant (*Sig.* < 10^{-15}). This confirms that the review time depends highly on transcription quality and, to a lesser extent, on the language considered. It is worth noting, however, that most data points (videos) are for Spanish (207 out of 277), and thus results are certainly biased towards this language. In this regard, a closer look at the distribution of data points reveals that they are more or less clustered by language. This was not unexpected since, after all, there are language- and MOOC-dependent factors (e.g., topic, reviewers and review quality requirements) that certainly have some effect on the RTF but fall out of the scope of this work. In any case, the statistical significance of the fit suffices to support the idea that RTF mainly depends on WER, irrespective of the transcription language. For example, and to be more precise, taking a couple of reference points on the logarithmic curve we can infer that a one-hour video transcription of 10 WER points will take 3 hours to be reviewed, and a video of the same duration with 20 WER points of transcription error will require almost 4 hours. This is significantly less time than the 10 RTF for transcribing from scratch.

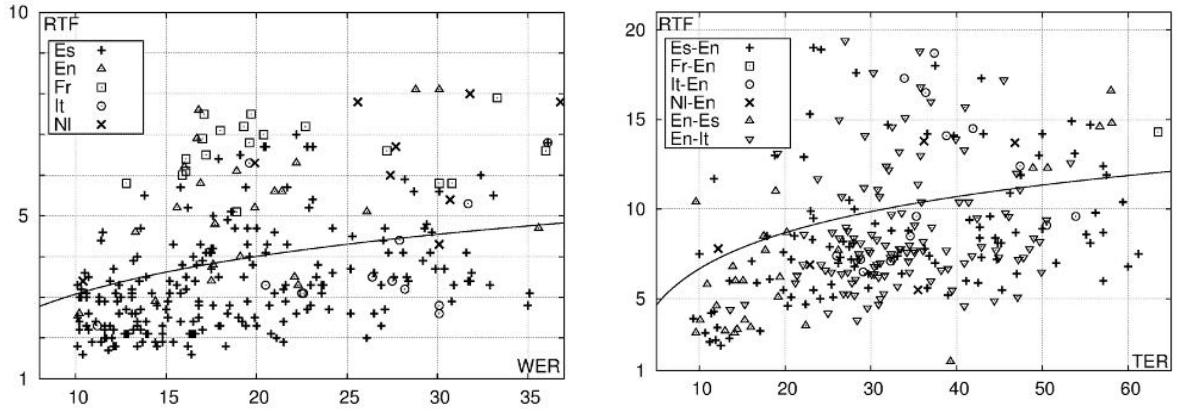

Figure 6. RTF vs. WER per transcription language (left) and RTF vs. TER per translation pair (right)

As with WER, the fitted logarithmic model shown at the bottom of Figure 6 (R^2 = 0.78, β = 2.90) is statistically significant (*Sig.* < 10^{-15}) for RTF prediction from TER. As above, then, we can confirm that RTF depends more on the translation quality (TER) than on the language pair considered. In contrast to the above results for WER, however, the distribution of data points does not reveal a clear language pair-dependent clustering structure. Taking into account that data points for Spanish (i.e., Spanish→English) are still dominant (250 out of 371), this adds more evidence to support the validity of the fitted logarithmic model. If, for example, a reviewed one-hour video transcription is automatically translated with about 30 TER points, then we may expect an RTF of around 9, that is, 9 hours for reviewing the translation. This is much less than the 30 hours (30 RTF) we may expect if translation is carried out manually from scratch; in other words, it entails a review time saving of 70% relative.

Impact on the case studies

Over the last two years, we have been collecting precise statistics on multilingual data consumption in the two real-life case studies mentioned above: the EMMA platform and the UPV media repository. This data is summarized below in order to better gauge the impact that the availability of video transcriptions and translations has had on both case studies.

The EMMA platform

Table 8 shows the number of native and non-native students enrolled in MOOCs offered on the EMMA platform, organized by the original language of the course. It goes without saying that non-native students could only follow these MOOCs thanks to the TLP-based multilingual component in EMMA described above. The last column in Table 8 shows the relative increase in the total number of students over native students due to the enrolment of non-native students.

Table 8. Statistics on student enrolment in MOOCs on the EMMA platform

Language	Native students	Non-native students	Relative increase (%)
Spanish	161	547	340
French	983	879	89
Italian	609	259	43
Dutch	501	104	21
English	351	27	8
Total	2605	1816	70

Note that the results in Table 8 are given in decreasing order respect to the relative increment of non-native students. The best results were obtained by MOOCs originally in Spanish and followed by 161 Spanish-speaking students. As these courses were also delivered in English and Italian, 547 non-Spanish-speaking students enrolled in the courses, increasing the total number of students by 340% with respect to the Spanish-speaking students. MOOCs in French almost doubled their number of students by offering these courses also in English. MOOCs in Italian and Dutch translated into English also experienced a relative increase with the non-native students enrolled of approximately 40% and 20%, respectively. Finally, English courses translated into Spanish had a small relative increase in student enrolment, mainly explained by the fact that English is considered a lingua franca and many non-native students are able to follow the course in English, at least students at this level of education. Overall, the translated versions of the MOOCs facilitated by the TLP in the EMMA platform attracted students that are non-native in the original language of the courses, increasing the total student enrolment by a notable 70%.

Indeed, according to exit questionnaires filled in by almost 1500 students enrolled in EMMA courses, 75% of them appreciated multilinguality as a feature of this platform and 70% found multilingual subtitles useful (Ferrari et al., 2016a). Taking into account only those approximately 200 students that replied to mini-questionnaires embedded in 17 running MOOCs, 31% of them always used the translation functionality, that is, the MOOC was originally in a different language from their mother tongue; and 29% of them sometimes used the translation functionality. Indeed, at least 90% of the students using always or sometimes the translation functionality agreed that this functionality enhances the overall value of the EMMA platform and makes EMMA a truly European experience (Ferrari et al., 2016b).

The UPV media repository

Table 9 shows the number of poliMedia videos and subtitle views (in thousands) per language and in total from June 2015, when view logs were activated, to May 2016.

Table 9. Video and subtitle views (in thousands) per language and total

Video language	Video views	Subtitle views	
		Spanish	English
Spanish	629	6.9	1.1
English	63	1.3	0.5
Total	692	8.2	1.6

The main conclusion that can be drawn from Table 9 is that, on average, subtitles were turned on in 1.4% of video views. It is worth noting, however, that a 1.4% of a large number of video views (i.e., almost 700K over the last year) is a significant number of users turning subtitles on (i.e., almost 10K over the last year). Indeed, in relative terms, it is interesting to observe that 2.5% of the English-language videos had their subtitles activated, in contrast to Spanish-language videos which did in 1.3% of the views. This result does not come as a surprise since most UPV students are native Spanish speakers with English as a foreign language. Finally, Spanish subtitles were predominant when subtitles were activated, being chosen in 86% and 71% of the cases for Spanish- and English-language videos, respectively.

Apart from the accessibility benefits for hearing-impaired and foreign students, the availability of transcriptions has allowed for the indexing and subsequent search for specific words in this large video lecture repository. Indeed, this search tool at the UPV media repository allows students to find the specific video clip in which a word is uttered by the lecturer. Thus, students can discard video clips that are not of their interest to focus on those ones in which a specific concept is explained, saving a significant amount of time. Subtitles also provide support for students in their arduous note-taking tasks.

Conclusions

In this work, we have reported a large part of the experience we have gained from producing low-cost multilingual video subtitles of publishable quality for MOOCs and OER. Apart from describing the systems, tools and integration components employed for such purpose, a comprehensive evaluation of the results achieved has been provided from three viewpoints: the quality of video transcriptions and translations automatically generated from task-adapted ASR/MT systems, the time required to review them, and the impact multilingual subtitles have had on a MOOC platform and a large video lecture repository.

The quality of automatic transcriptions and translations has been proved to be in most cases below 25% of WER and 50% of TER, respectively. This means that it is worth post-editing them to achieve publishable subtitles instead of generating them ex novo. Indeed, the output of the adapted ASR/MT systems has been positively compared to state-of-the-art automatic transcription and translation tools provided by mainstream providers. More precisely, these systems are on average 38% and 17% better than YouTube's automatic captioning and Google Translate, respectively.

Regarding the review process, we have showed that a lecturer can save between 30% and 70% of the time devoted to review transcriptions, and between 25% and 75% of the translation review time, with respect to performing these tasks from scratch. In addition, a multilingual linear regression model has been proposed to infer the review time (RTF) as a function of WER in the case of transcription, and in terms of TER for translation.

The availability of multilingual video subtitles has been shown to have a great impact in our case studies. On the one hand, in the EMMA platform, the translation of MOOCs into a second, or even a third language has significantly increased course visibility boosting student enrolment by 70% relative. On the other hand, multilingual subtitles at the UPV media repository have not only improved accessibility to the video lectures for hearing-impaired and non-native-speaking students, but also have allowed the development of added-value functionalities such as indexing and search capabilities, and obviously translated subtitles.

Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287755 (transLectures) and from the EU's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme under grant agreement no. 621030 (EMMA). Additionally, it is supported by the Spanish research project TIN2015-68326-R (MINECO/FEDER).

References

del-Agua, M. A., Giménez, A., Serrano, N., Andrés-Ferrer, J., Civera, J., Sanchis, A., & Juan, A. (2014). The transLectures-UPV toolkit. In J. L. Navarro-Mesa et al. (Eds.), *Proceedings of IberSpeech* (pp. 269–278). Las Palmas, Spain: Springer, Heidelberg.

Akita, Y., Mimura, M., & Kawahara, T. (2009). Automatic transcription system for meetings of the Japanese national congress. In M. Uther et al. (Eds.), *Proceedings of Interspeech* (pp. 84–87). Brighton, UK: International Speech Communication Association.

Axelrod, A., He, X., & Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of Empirical Methods on Natural Language Processing* (pp. 355–362). Edinburgh, UK: Association for Computational Linguistic.

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication, 33*(1–2), 5–22.

Bishop, J., & Verleger, M. A. (2013). The Flipped classroom: A Survey of the research. In *Proceedings of American Society for Engineering Education* (Vol. 30, No. 9, pp. 1-18). Atlanta, GA: American Society for Engineering Education.

Ferrari, C., Pennati, C., Tontodonati, A., Tammets, K., Panto, E., Marcellin, L., & Politi, R. (2016a, July). *D4.3.2 data and impact analysis report*. Deliverable of the European EMMA project.

Ferrari, C., Pennati, C., Tontodonati, A., Tammets, K., Panto, E., Marcellin, L., & Politi, R. (2016b, July). *D4.4 pilot cycle evaluations*. Deliverable of the European EMMA project.

Hazen, T. J. (2006). Automatic alignment and error correction of human generated transcripts for long speech recordings. In R. M. Stern (Ed.), *Proceedings of Interspeech 2006* (pp. 1606–1609). Pittsburgh, PA: International Speech Communication Association.

Hunt, M. J. (1990). Figures of merit for assessing connected-word recognisers. *Speech Communication, 9*(4), 329–336.

Kay, R. H. (2012). Exploring the use of video podcasts in education: A Comprehensive review of the literature. *Computers in Human Behavior, 28*(3), 820–831.

Ketterl, M., Schulte, O. A., & Hochman, A. (2010). Opencast matterhorn: A Community-driven open source software project for producing, managing, and distributing academic video. *Interactive Technology and Smart Education, 7*(3), 168–180.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of Association for Computational Linguistics* (pp. 177–180). Prague, Czech Republic: Association for Computational Linguistics.

Kordoni, V., Bosch, A., Kermanidis, K., Sosoni, V., Cholakov, K., Hendrickx, I., Huck, M., & Way, A. (2016). Enhancing access to online education: Quality machine translation of MOOC content. In N. Calzolari et al. (Eds.), *Proceedings of Language Resources and Evaluation* (pp. 16–22). Paris, France: European Language Resources Association.

Martínez-Villaronga, A., del-Agua, M., Andrés-Ferrer, J., & Juan, A. (2013). Language model adaptation for video lectures transcription. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing* (pp. 8450–8454). Vancouver, Canada: IEEE.

Munteanu, C., Baecker, R., Penn, G., Toms, E., & James, D. (2006). The Effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In R. Grinter et al. (Eds.), *Proceedings of the Conference on Human Factors in Computing Systems* (pp. 493–502). Montreal, Canada: ACM.

O'Brien, S., & Simard, M. (2014). Introduction to special issue on post-editing. *Machine Translation, 28*(3-4), 159-164.

de-Pablos, P. O., Zhao, J., & Tennyson, R. (Eds.) (2011). *Technology enhanced learning for people with disabilities: Approaches and applications*. Hershey, PA: IGI Global.

Plitt, M., & Masselot, F. (2010). A Productivity test of statistical machine translation postediting in a typical localisation context. *Prague Bulletin of Mathematical Linguistics, 93*, 7–16.

Ranchal, R., Taber-Doughty, T., Guo, Y., Bain, K., Martin, H., Robinson, J. P., & Duerstock, B. S. (2013). Using speech recognition for real-time captioning and lecture transcription in the classroom. *IEEE Transactions on Learning Technologies, 6*(4), 299–311.

Reidsma, D., Hofs, D., & Jovanovic, N. (2005). Designing focused and efficient annotation tools. In L. P. J. J. Noldus et al. (Eds.), *Proceedings of Measuring Behaviour* (pp. 149–152). Wageningen, The Netherlands: Noldus Information Technology.

Repp, S., Groß, A., & Meinel, C. (2008). Browsing within lecture videos based on the chain index of speech transcription. *IEEE Transactions on Learning Technologies, 1*, 145–156.

Shah, D. (2015). *By the numbers: MOOCS in 2015*. Retrieved from http://www.class-central.com/report/moocs-2015-stats

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas* (pp. 223–231), Cambridge, MA: Association for Machine Translation in the Americas.

Specia, L. (2011). Exploiting objective annotations for measuring translation post-editing effort. In M. L. Forcada et al. (Eds.), *Proceedings of European Association for Machine Translation* (pp. 73–80). Leuven, Belgium: European Association for Machine Translation.

Specia, L., Turchi, M., Wang, Z., Shawe-Taylor, J., & Saunders, C. (2009). Improving the confidence of machine translation quality estimates. In *Machine Translation Summit XII* (pp. 136-143). Ottawa, Canada: Association for Machine Translation in the Americas.

Turró, C., Ferrando, M., Busquets, J., & Cañero, A. (2009). Polimedia: A System for successful video e-learning. In *European University Information Systems 2009*. Santiago de Compostela, Spain: European University Information Systems.

Valor-Miró, J. D., Silvestre-Cerdà, J. A., Civera, J., Turró, C., & Juan, A. (2015). Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories. *Speech Communication, 74*(C), 65–75.

Williams, J. D., Melamed, I. D., Alonso, T., Hollister, B., & Wilpon, J. (2011). Crowd-sourcing for difficult transcription of speech. In *Proceedings of Automatic Speech Recognition and Understanding* (pp. 535–540). doi:10.1109/ASRU.2011.6163988