A System Architecture to Support Cost-Effective Transcription and Translation of Large Video Lecture Repositories

Joan Albert Silvestre-Cerdà, Alejandro Pérez, Manuel Jiménez,
Carlos Turró, Alfons Juan and Jorge Civera.
DSIC/ITI/ASIC, Universitat Politècnica de València, València, Spain
{jsilvestre, aperez, ajuan, jcivera}@dsic.upv.es, jimenez@asic.upv.es, turro@cc.upv.es

***Abstract*—Online video lecture repositories are rapidly growing and becoming established as fundamental knowledge assets. However, most lectures are neither transcribed nor translated because of the lack of cost-effective solutions that can give accurate enough results. In this paper, we describe a system architecture that supports the cost-effective transcription and translation of large video lecture repositories. This architecture has been adopted in the EU project transLectures and is now being tested on a repository of more than 9000 video lectures at the Universitat Politècnica de València. Following a brief description of this repository and of the transLectures project, we describe the proposed system architecture in detail. We also report empirical results on the quality of the transcriptions and translations currently being maintained and steadily improved.**

***Index Terms*—Language Technologies, Machine Translation, Automatic Speech Recognition, Massive Adaptation, Intelligent Interaction, Education, Video Lectures, Multilingualism, Accessibility, Opencast Matterhorn**

I. Introduction

Online multimedia repositories are rapidly growing and becoming established as fundamental knowledge assets. This is particularly true in the area of education, where large repositories of video lectures are being built on the back of on increasingly available and standardised infrastructure [1], [2]. However, most of these lectures are neither transcribed nor translated due to a lack of cost-effective solutions to do so in a way that gives accurate enough results. Solutions of this kind are clearly necessary in order to make these lectures accessible to speakers of different languages and to people with hearing disabilities [3]. They would also facilitate lecture searchability and analysis functions, such as classification, summarisation [4] and plagiarism detection.

In this paper, we describe a system architecture that can support the cost-effective transcription and translation of large video lecture repositories. This architecture has been adopted in the EU project transLectures, whose main aim is to achieve just this through the use of advanced automatic speech recognition (ASR) and machine translation (MT) technologies. The starting hypothesis in transLectures is that the gap that must be bridged by these technologies in order to achieve acceptable results for the kind of audiovisual collections being considered is relatively small. In transLectures, two key lines of research are being pursued: *massive adaptation* and *intelligent (user) interaction* [5].

Massive adaptation refers to process of exploiting the wealth of knowledge available about these video lecture repositories (lecture-specific knowledge, such as speaker, topic and slides) to create a specialised, “in-domain” transcription and translation system. A system adapted using this knowledge is therefore likely to produce a far better ASR and MT output than a general-purpose system.

Intelligent interaction is the process of human-computer interaction whereby we can exploit feedback from the user or “prosumer” community of a given video lecture repository. For instance, we can more or less count on a university lecturer being willing to devote a few minutes of his time to correcting any errors in the automatic transcript generated for a lecture he has recently recorded. This is the kind of interaction we are exploring with poliMedia. The intelligent part comes in when deciding exactly which segments of the transcript the lecturer should be asked to interact with. For example, the system should not simply present the first minute of a lecture for review, since this section may well be perfectly transcribed and need no manual corrections. This would be a waste of user effort. Instead, an intelligent system should first identify which section(s) of the lecture contain the most errors; that is, which section(s), based on its automatic confidence measures, are most likely to contain errors, and then present these sections only to the the user for correction. The system can then use these corrections to re-train the underlying models and thereby avoid the same errors occurring in future transcriptions or translations.

The above ideas are being tested on two case studies: VideoLectures.NET [1] and poliMedia [6]. VideoLectures.NET is an online video repository with more then 14,000 talks (10,000 hours) given by top researchers in various academic settings. poliMedia is a collection of over 9,000 videos (2,000 hours) recorded by course lecturers under controlled conditions at the Universitat Politècnica de València, Valencia (Spain). Both repositories are active players in the diffusion of the open-source Matterhorn platform [7] currently being adopted by many education institutions and organisations within the Opencast community. Indeed, a third key premise of transLectures is to use (and develop) a system architecture that works with Matterhorn, to allow the rapid adoption and real-life testing of transLectures technologies.

In this paper, we aim to describe the transLectures system architecture and present the first results achieved in the context of poliMedia. First, we give an overview of related work, in Sec. II. Then, following a brief description of poliMedia (Sec. III) and transLectures (Sec. IV), we describe the proposed architecture in detail (Sec. V). In Sec. VI, we report empirical results on the quality of the transcriptions and translations currently being maintained and steadily improved. Finally, we draw some key conclusions in Sec. VII.

II. Related Work

The sheer volume of data that has to be processed in large multimedia repositories means that automatic data processing is a truly challenging task [8]. This is particularly true in the areas of automatic speech recognition (ASR) and statistical machine translation (SMT) given the computational complexity of both tasks.

The integration of ASR technologies into multimedia repositories is a well-known problem which has been previously and successfully explored, especially in the case of news broadcasting [9], [10] and TV content in general [11], although most of these systems are mainly designed to provide subtitles in real-time. Other ASR applications can be found in the subtitling of Parliament sessions [12] where, in some cases, manual transcripts are synchronized with the input audio signal in order to generate the corresponding subtitles [13]; also in video indexing [14].

The integration of SMT systems, meanwhile, has not been explored in any great depth. There is one exception to this [15], where manual transcripts are assumed to be available before the translation process starts.

No previous works have explored the integration of both ASR and SMT technologies into large media repositories, nor the deployment of a platform through which users can amend errors in recognition and/or translation with little effort.

III. poliMedia

poliMedia is a recent, innovative service for the creation and distribution of multimedia educational content at the UPV. It is designed primarily to allow UPV professors to record their courses in videos lasting around 10 minutes and accompanied by time-aligned slides. It serves more than 36,000 students and 2800 university lecturers and researchers. poliMedia began in 2007 and has already been exported to several universities in Spain and South America. Table I shows basic statistics of the poliMedia repository.

TABLE I
Basic statistics of the poliMedia repository.

Lectures	9222
Duration (hours)	2102
Avg. Lecture Length (minutes)	13
Speakers	1302
Avg. Lectures per Speaker	7

poliMedia recordings are made up of two videos stacked horizontally: one of the slides and another of the speaker, with a resolution of 1280x720 points. See Figure 1 for an example of a poliMedia recording.

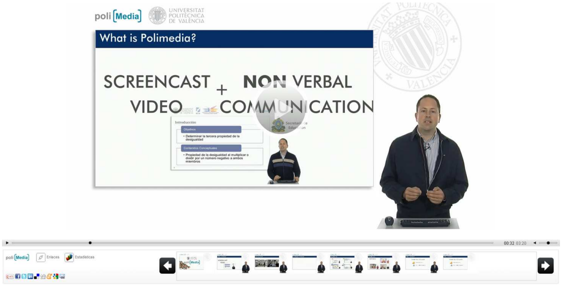


Fig. 1. Example of a poliMedia lecture.

The production process for a poliMedia repository has been carefully designed to achieve both a high rate of production and a high quality output, comparable to that of a TV production but at a lower cost. A poliMedia production studio consists of a 4x4 metre room with a white background in which we find a video camera; a capture station; a pocket microphone; lighting; and A/V equipment including a video mixer and an audio noise gate. It is worth noting that the use of lighting in such a small set allows us to get a sharper image much more easily than in a lecture recording hall. The hardware cost of this studio stands at around 15,000 euros.

The recording process is quite simple: lecturers (speakers) are invited to come to the studio with their presentation and slides. They deliver their lecture, while they and their computer screen are recorded in two different streams. The two streams are put side-by-side to generate a raw preview of the poliMedia content, which can be reviewed by the speaker at any time.

If the speaker is satisfied with the end result, a post-process is applied to the raw recordings, which includes cropping, joining (with a little overlap) and h.264 encoding, in order to generate a mp4 file suitable for distribution. This process is fully automatic, meaning that the speaker can review the post-processed video file in just a few minutes. Next, the mp4 file is distributed online via a streaming server.

Please visit http://polimedia.blogs.upv.es/?lang=en for more details and examples[1].

IV. transLectures

transLectures (acronym of Transcription and Translation of Video Lectures) is an EU (FP7-ICT-2011-7) STREP project in which advanced automatic speech recognition and machine translation techniques are being tested on large video lecture repositories. The project began in November 2011 and will run for three years. The transLectures consortium includes video lecture providers (users), experts in ASR and MT, and professional transcription and translation providers:

- *UPV*: Universitat Politècnica de València, Valencia, Spain.

[1] See http://polimedia.upv.es/visor/?id=39f62a9a-4cf5-bd4e-92f3-cb34e4792a85 for a brief presentation in Spanish, with subtitles available in English.

- *XEROX*: Xerox S.A.S., Grenoble, France.
- *JSI*: Jozef Stefan Institute, Ljubljana, Slovenia.
- *RWTH*: Rheinisch-Westfaelische Technische Hochschule, Aachen, Germany.
- *EML*: European Media Laboratory GmbH, Heidelberg, Germany.
- *DDS*: Deluxe Digital Studios Ltd, London, UK.

The UPV coodinates the transLectures project, with Alfons Juan-Ciscar as project coordinator. transLectures is grounded on the three following scientific and technological objectives:

1) *Improvement of transcription and translation quality by massive adaptation.*
   ASR has not yet revealed its full potential in the generation of acceptable transcriptions for large-scale collections of audiovisual objects. However, that potential is there and relatively little further research into ASR technology is required; rather we must learn to better exploit the wealth of knowledge we have at hand. More precisely, it is our aim to demonstrate that acceptable transcriptions can be obtained through the massive adaptation of general-purpose models from lecture-specific knowledge such as speaker, topic and, more importantly, time-aligned slides.
2) *Improvement of transcription and translation quality by intelligent interaction.*
   Massive adaptation can deliver substantial contributions to the improvement of overall quality, but it is our belief that sufficiently accurate results are unlikely to be obtained through fully-automated approaches alone. Instead, in order to reach the desired levels of accuracy, we must consider user interaction. Current user models for the transcription and translation of audiovisual objects are batch-oriented. Under this model, an initial transcription/translation is first computed by the system offline and then sent to the user to be post-edited manually without system participation. In our view, these models only yield satisfying results when highly collaborative users are working on near-perfect system output. Otherwise, a more intelligent interaction model is required that saves on user supervision and allows the system to learn from user supervision actions. In transLectures, our aim is to develop innovative, truly interactive models in which the system learns from, and reacts to, each user supervision action immediately.
3) *Integration into Matterhorn to enable real-life evaluation.*
   In contrast to many past research efforts in which system prototypes are evaluated in the lab alone and are largely inapplicable to real-life settings, we will be developing tools and models for use with Matterhorn. We will therefore be able to evaluate their usefulness using real-life data in a real-life context.

We are convinced that, upon successful achievement of our objectives, our innovative solutions will be rapidly deployed across many educational repositories in Europe and worldwide, enabling them to overcome language barriers and reach wider audiences while supporting linguistic diversity.

At transLectures, we are testing our ideas on VideoLectures.NET and on poliMedia, which is also part of the Matterhorn Community. For automatic transcription, we are considering English and Slovenian in VideoLectures.NET (which account for more than 90% of all video lectures), and Spanish in poliMedia. Meanwhile, automatic translation is carried out from {Spanish, Slovenian} into English, and English into {French, German, Slovenian, Spanish}.

V. System Architecture

This section describes the architecture of the transLectures system deployed in the poliMedia repository. Figure 2 shows a global overview of the components and processes involved in the transLectures system.

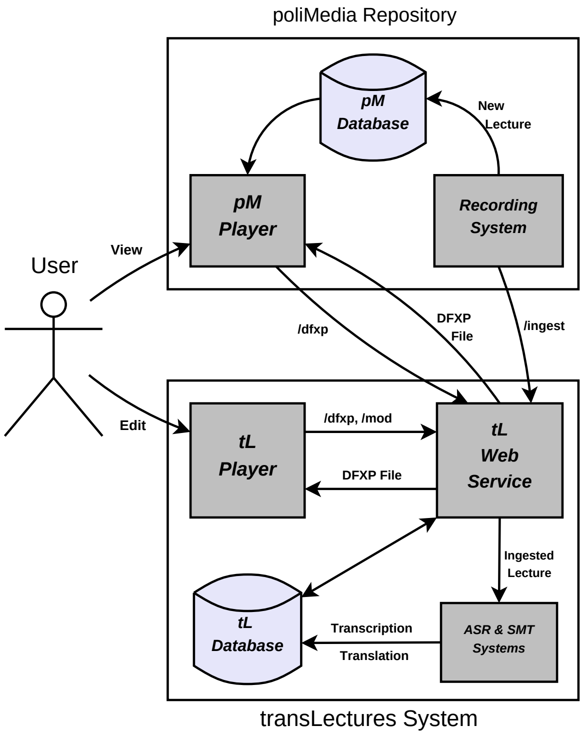


Fig. 2. Overview of the transLectures system in the poliMedia repository.

First of all, we have identified two use cases to cover the different ways in which the user is able to interact with the system. Specifically, the user may want simply to view a video lecture and its corresponding captions, which would be the first use case, or he might choose to also edit/supervise the automatic captions, which would be the second.

In the first use case, the user browses the poliMedia catalogue and selects the lecture he wants to watch. The user then is redirected to the poliMedia player (Figure 1), where he can watch the requested lecture. This player allows the user to select captions for the video in different languages, where available. The list of languages available is obtained by sending a request to an external service: the transLectures web service (see Section V-A). This web service checks to see if there are captions available for a given lecture and, if so, in which languages. Once the user has selected a language, the player sends a request to the web service asking for the corresponding caption file. The caption file is then sent in DFXP format [16] and immediately displayed on the player. All captions, along with the corresponding metadata, are stored in the transLectures database (see Section V-C).

In the second use case, the user, while watching a lecture with the corresponding captions as described in the first use case, notices that the captions displayed contain transcription errors and decides to correct them. The poliMedia player includes an *Edit* button which, once pressed, redirects the user to the transLectures player. This player has caption editing capabilities, among other features (see Section V-B). Corrections made by the user are sent back to the web service, appending them into the original DFXP file. The DFXP format is extended in this use case in order to be able to track the history of modifications made by users and the automatic systems, allowing the player to show the best captions available for every segment. In other words, a DFXP file can be understood as a mini repository of caption modifications. Lastly, the corrections made by the user are committed to the transLectures database and immediately propagated to the ASR and statistical MT (SMT) systems (see Section V-D), in order to improve the underlying models on the basis of this user feedback.

The transLectures system needs to be permanently synchronised with the poliMedia repository in order to provide transcriptions and translations for any newly recorded videos. For this purpose, the transLectures web service provides a lecture upload service, known as *ingest*, which is used by the poliMedia recording system. Then, once a new lecture has been uploaded to the transLectures web service via the ingest interface, the transcription of this lecture, and its subsequent translation into different languages, is carried out by the ASR and SMT systems.

It is worth noting the distributed nature of the transLectures system architecture (i.e. that each component could be deployed on a different machine), although in this case study all components are hosted by a single machine mounting an Intel Core i7-3820 CPU @ 3.60GHz with 64GB of RAM.

In the following sections, we give more detailed descriptions of the key components of the transLectures system.

A. transLectures web service

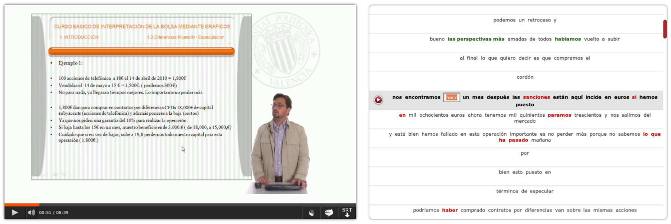
The transLectures web service is the interface for exchanging information and data between the poliMedia repository and the transLectures system. It also enables the subtitle visualisation and editing capabilities of the transLectures player. This web service is implemented as a python Web Server Gateway Interface (WSGI), and defines a set of HTTP interfaces related to caption delivery and media upload:

- *ingest*: POST request which allows the client to upload audio/video files and other related material, such as slides and textual resources, which could be useful for adapting the ASR system. This POST request also allows additional metadata to be submitted, such as the language of the recording and speaker ID. This metadata can be used to enhance the accuracy of the ASR system by applying speaker adaptation techniques. The automatic translation is then generated from the automatic transcription, and both are stored in DFXP format.
- *status*: GET request to check the status of a video lecture uploaded through the ingest interface.
- *langs*: GET request that provides the client with a list of the captions and languages available for a given lecture.
- *dfxp*: GET request that returns the captions in DFXP format for a specific lecture and language.
- *mod*: POST request that sends and commits changes made by a user when supervising a transcription or translation. These changes are also sent to the ASR and SMT systems in order to take advantage of user corrections and update the underlying models.

All these interfaces operate with the transLectures database, which stores all information needed by the transLectures web service.

B. transLectures Player

A PHP/HTML5 video player and caption editor has been carefully designed to expedite the error supervision task and obtain cost-effective captions of an acceptable quality in exchange for a minimum amount of user effort.



CURSO BÁSICO DE INTERPRETACIÓN DE LA BOLSA MEDIANTE GRÁFICOS

1. INTRODUCCIÓN 1.3 Diferencias Inversión - Especulación.

- Ejemplo 1:
- 100 acciones de telefónica a 18€ el 14 de abril de 2010 = 1,800€
- Vendidas el 14 de mayo a 15 € = 1,500€. (perdemos 300 €)
- No pasa nada, ya llegaran tiempos mejores. Lo importante no perder más.
- 1,800€ dan para comprar en contratos por diferencias CFDs 18,000€ de capital subyacente (acciones de telefónica) y además ponerse a la baja (cortos)
- Ya que nos piden una garantía del 10% para realizar la operación.
- Si baja hasta los 15€ en un mes, nuestro beneficio es de 3.000.€ (de 18,000, a 15,000,€)
- Cuidado que si en vez de bajar, sube a 19,8 perdemos todo nuestro capital para esta operación (1.800€.)

podemos un retroceso y

bueno las perspectivas más amadas de todos habíamos vuelto a subir

al final lo que quiero decir es que compramos el

cordón

nos encontramos hace un mes después las sanciones están aquí incide en euros si hemos puesto

en mil ochocientos euros ahora tenemos mil quinientos paramos trescientos y nos salimos del mercado

y está bien hemos fallado en esta operación importante es no perder más porque no sabemos lo que ha pasado mañana

por

bien esto puesto en

términos de especular

podríamos haber comprado contratos por diferencias van sobre las mismas acciones

Fig. 3. transLectures Player showing the default side-by-side editing layout. The video playback is shown on the left-hand side of the screen, whilst the transcription editor appears on the right-hand side. In the figure above, the player is displaying the intelligent interaction mode in which the system asks the user to amend only those words considered most likely to be incorrect by the system (highlighted in red).

Three alternative editing layouts are available for users to choose from according to their personal preferences. Figure 3 shows the default side-by-side editing interface with the video playback on the left and transcription editor on the right. Currently, the transLectures player offers two interaction modes: batch interaction, in which the user can freely supervise any segment of the video; and intelligent interaction, in which the player asks the user to correct only those words considered most likely to be incorrect by the system. Additionally, a complete set of key shortcuts has been implemented to enhance expert user capabilities. Other helpful features are being continuously added to the editor in response to the user feedback currently being gathered.

Captions are retrieved from the transLectures web service through the *dfxp* interface. Once retrieved, users are able to correct transcription/translation errors and savee modifications back to the web service through its *mod* interface.

C. transLectures database

The transLectures database is at the core of the transLectures system. It is a PostgreSQL relational database which stores all the data required by the transLectures web service, and the ASR and SMT systems. Specifically, the transLectures database stores the following entities:

- Lectures: All the information related to a specific lecture is stored in the database, such as language, duration, title, keywords and category. In addition, an external ID, recognised by the poliMedia database, is stored and used in all transactions performed between the poliMedia player and the transLectures web service for lecture identification purposes.
- Speakers: The information about the speaker of a given lecture can be exploited by the ASR system, adapting the underlying models to the speech peculiarities of a specific speaker. The result is a better transcription and, consequently, a better translation.
- Captions: All captions generated by the ASR and SMT systems are kept in the database and retrieved by the transLectures web service.
- Uploads: Every time a web service ingest operation is requested by the poliMedia recording system, a new entry is stored in the database. Once the computation of the automatic transcription and translation of the lecture is performed, a new entry is added to the lectures, speakers and captions entities.

D. ASR & SMT Systems

ASR and SMT systems are key components of the transLectures system. These systems generate the automatic transcriptions and translations of every lecture available in the poliMedia repository.

Our ASR system is made up of the transLectures-UPV open source toolkit, TLK [17], which consists of a set of tools that allow acoustic model training and the recognition of audio signals such as video lectures. It provides features similar to those of other Besides, the SRILM toolkit [18] is used to deploy n-gram language models. The two main components of an ASR system are acoustic models and language models. Our ASR system is designed to exploit all available information regarding the poliMedia repository, in order to enhance transcription quality. Information about the speaker is used to inform acoustic model adaptation techniques, while text extracted from slides and other related textual resources are used to adapt language models to the specific topic of the lecture.

Meanwhile, our SMT system is based on the well-known open source Moses toolkit [19]. The translations of a given lecture into different languages are generated from its automatic transcription. This means that translation accuracy is highly correlated with transcription quality; that is, the better the transcription, the better the translation.

Transcriptions and translations are automatically regenerated following major upgrades to the ASR or SMT system, meaning that the repository's overall transcription and translation quality is constantly improving. These upgrades might be the result of better acoustic, translation or language models, or of new ASR and SMT techniques.

Further details about the training of ASR and SMT systems and their performance are described in Section VI.

VI. System Evaluation

The transLectures system deployed in the poliMedia repository is being evaluated from two different viewpoints. Firstly, transcription and translation quality are constantly assessed to gauge the progress of the underlying ASR and SMT systems. Secondly, user satisfaction and productivity are analysed through internal and external evaluations.

The quality of the ASR and MT systems is monitored using automatic measures such as *Word Error Rate* (WER) for transcription, and *Bilingual Evaluation Understudy* (BLEU) [20] and *Translation Error Rate* (TER) [21] for translation. Automatic measures allow the easy evaluation of system performance during development, without the expensive intervention of transcription or translation experts.

To this end, 114 hours of Spanish poliMedia video lectures were manually transcribed and approximately 15 hours were translated into English. This manually transcribed and translated data was partitioned into training, development and test sets that were allocated to training, tuning and evaluating our ASR and MT systems. Statistics on these three sets are shown in Table II.

TABLE II
Statistics on the training, development and test sets allocated to automatic evaluations in poliMedia.

	Training	Development	Test
Videos	655	26	23
Speakers	73	5	5
Hours	107h	3.8h	3.4h
Sentences	39.2K	1.3K	1.1K
Words	936K	35K	31K
Vocabulary	26.9K	4.7K	4.3K

Table III reports the evolution of transcription quality as additional resources and adaptation techniques are incorporated into the baseline Spanish ASR system. As mentioned above, transcription quality is expressed in terms of WER, which can be loosely understood as the percentage of words that need to be amended in some way in order to obtain a correct transcription.

TABLE III
Evolution of transcription quality for the Spanish ASR system in poliMedia.

	WER
Baseline	36.0
+External resources	30.3
+CMLLR	24.6
+LMslides	22.6

The baseline system scores 36.0 WER points. This is significantly improved down to 30.3 WER points when more training data from external linguistic resources is used to train the ASR system. The application of conventional speaker adaptation techniques, such as CMLLR [22], leads to considerable improvement of transcription quality, decreasing WER to 24.6 points. Finally, the adaptation of the language models used in the ASR system using the text content extracted from the accompanying presentation slides gives the best result: 22.6 WER points.

Manual Spanish into English (ES-EN) translations for poliMedia are limited to the development and test sets, since only 7 hours were translated. Training data was therefore collected from publicly available parallel corpora. Adaptation techniques were applied based on information available about lecture topic.

Table IV summarises the baseline and topic adaptation results obtained for Spanish into English translations in terms of BLEU and TER. BLEU scores can be intuitively understood as the degree of overlap between the automatic translation generated by the MT system and the reference translation provided by a professional linguist. TER, meanwhile, can be thought of as a percentage approximation of the number of words that need to be corrected in order to achieve the reference translation. However, both measures of translation quality are pessimistic, since the use of a single reference translation ignores the possibility of other correct translations that might be more similar to the automatic translation.

TABLE IV
BLEU AND TER RESULTS ON THE POLIMEDIA CORPUS FOR THE ES-EN SMT SYSTEM.

	BLEU	TER
Baseline	23.4	56.5
+Topic adaptation	26.0	54.6

The ES-EN baseline system obtained 23.4 BLEU and 56.5 TER points, which were drastically improved by applying adaptation techniques based on a selection of topic-specific sentences related to the topic of the videos being translated.

Internal evaluation trials with UPV lecturers are currently in progress, though the initial phase has been successfully completed. A series of refinements were made to the player interface in response to user feedback and a usability survey revealed a high level of satisfaction with the current system. Furthermore, user interaction statistics are being collected and analysed to develop a user model that will be exploited during the development of intelligent interaction techniques designed to minimise user effort.

VII. Conclusions

In this paper, we have presented our system for allowing online video lecture repositories to provide users acceptable transcriptions and translations in exchange for relatively little user effort. It should be noted that the transLectures system is based entirely on open-source, freely available software. We have also described the transLectures system's main components; namely, the transLectures web service, player and database, and the ASR and SMT systems, as well as how these components interact with each other under two use cases: viewing and editing video lectures. Evaluations of the system using conventional metrics report good scores, indicating that the transcriptions and translations obtained are of an acceptable quality. Moreover, real-life that users have rated these captions positively in the initial phases of an internal evaluation.

Further developments and improvements on the transLectures system will focus on satisfying user needs as the transLectures player functionality is enriched to minimise user effort on caption supervision.

Acknowledgment

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287755. Funding was also provided by the Spanish Government with the FPU scholarship AP2010-4349.

References

[1] "VideoLectures.NET: Exchange ideas and share knowledge," http://www.videolectures.net.

[2] "Coursera: Take the World's Best Courses, Online, For Free," http://www.coursera.org.

[3] M. Wald, "Creating accessible educational multimedia through editing automatic speech recognition captioning in real time," *Interactive Technology and Smart Education*, vol. 3, no. 2, pp. 131–141, 2006.

[4] J. Glass *et al.*, "Recent progress in the mit spoken lecture processing project," in *Proc. of INTERSPEECH*, 2007, pp. 2553–2556.

[5] UPVLC, XEROX, JSI-K4A, RWTH, EML, and DDS, "Transcription and Translation of Video Lectures," in *Proc. of EAMT*, 2012.

[6] "poliMedia: Video lectures from the Universitat Politècnica de València," http://polimedia.upv.es/catalogo.

[7] M. Ketterl *et al.*, "Opencast matterhorn: A community-driven open source solution for creation, management and distribution of audio and video in academia," in *Proc. of IEEE ISM*, 2009, pp. 687–692.

[8] X. Sevillano *et al.*, "Indexing large online multimedia repositories using semantic expansion and visual analysis," *MultiMedia, IEEE*, vol. 19, no. 3, pp. 53–61, 2012.

[9] J. Neto *et al.*, "Broadcast news subtitling system in portuguese," in *Proc. of ICASSP*, 2008, pp. 1561–1564.

[10] A. Ortega *et al.*, "Real-time live broadcast news subtitling system for spanish," in *Proc. of INTERSPEECH*, 2009, pp. 2095–2098.

[11] A. lvarez *et al.*, "Apyca: Towards the automatic subtitling of television content in spanish," in *Proc. of IMCSIT*, 2010, pp. 567–574.

[12] A. Pražák *et al.*, "Automatic online subtitling of the czech parliament meetings," in *Proc. of TSD*, 2006, pp. 501–508.

[13] G. Bordel *et al.*, "Automatic subtitling of the basque parliament plenary sessions videos," in *Proc. of INTERSPEECH*, 2011, pp. 1613–1616.

[14] S. Repp *et al.*, "Browsing within lecture videos based on the chain index of speech transcription," *Learning Technologies, IEEE Transactions on*, vol. 1, no. 3, pp. 145–156, 2008.

[15] M. Melero *et al.*, "Automatic multilingual subtitling in the etitle project," *Proc. of Translating and the Computer*, vol. 28, 2006.

[16] World Wide Web Consortium (W3C), "Distribution Format Exchange Profile (DFXP)," http://www.w3.org/tr/2006/cr-ttaf1-dfxp-20061116.

[17] The TransLectures-UPV team, "The TransLectures UPV toolkit (TLK)," http://www.translectures.eu/tlk.

[18] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. of ICSLP*, 2002.

[19] H. Hoang and P. Koehn, "Design of the Moses Decoder for Statistical Machine Translation," in *Proc. of ACL*, 2008.

[20] K. Papineni *et al.*, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proc. of ACL*, 2002, pp. 311–318.

[21] M. Snover *et al.*, "A Study of Translation Error Rate with Targeted Human Annotation," in *Proc. of AMTA*, 2006.

[22] D. Giuliani *et al.*, "Speaker normalization through constrained MLLR based transforms," in *Proc. of INTERSPEECH*, 2004.