Document downloaded from:

http://hdl.handle.net/10251/55925

This paper must be cited as:

Valor Miró, JD.; Spencer, RN.; Pérez González De Martos, AM.; Garcés Díaz-Munío, GV.; Turró Ribalta, C.; Civera Saiz, J.; Juan Císcar, A. (2014). Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures. Open Learning: The Journal of Open and Distance Learning. 29(1):72-85. doi:10.1080/02680513.2014.909722.

The final publication is available at

http://dx.doi.org/10.1080/02680513.2014.909722

Copyright Taylor & Francis (Routledge): SSH Titles

Additional Information

Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures

J. D. Valor Miró, R. N. Spencer, A. Pérez González de Martos, G. Garcés Díaz-Munío, C. Turró, J. Civera and A. Juan

Department of Information Systems and Computation, Universitat Politècnica de València, Valencia, Spain

Video lectures are fast becoming an everyday educational resource in higher education. They are being incorporated into existing university curricula around the world, while also emerging as a key component of the open education movement. In 2007 the Universitat Politècnica de València (UPV) implemented its poliMedia lecture capture system for the creation and publication of quality educational video content and now has a collection of over 10,000 video objects. In 2011 it embarked on the EU-subsidised transLectures project to add automatic subtitles to these videos in both Spanish and other languages. By doing so, it allows access to their educational content by non-native speakers and the deaf and hard-of-hearing, as well as enabling advanced repository management functions. In this paper, following a short introduction to poliMedia, transLectures and *Docència en Xarxa* (Teaching Online), the UPV's action plan to boost the use of digital resources at the university, we will discuss the three-stage evaluation process carried out with the collaboration of UPV lecturers to find the best interaction protocol for the task of post-editing automatic subtitles.

Keywords: video lectures; online learning; modes of interaction; automatic speech recognition; user evaluations

Introduction

Online video lecture repositories are fast becoming a staple feature of the Internet and an everyday educational resource in higher education. Used to supplement traditional lectures, they are being incorporated into existing university curricula around the world, to enthusiastic responses from students (Soong, Chan, Cheers, & Hu, 2006). In 2007 the Universitat Politècnica de València (UPV) implemented its poliMedia lecture capture system for the cost-effective creation and publication of quality educational video content (poliMedia, 2007). It now has a collection of over 10,000 video objects created by 1373 lecturers, in part incentivised by the *Docència en Xarxa* (Teaching Online) action plan to boost the use of digital resources at the University. It has also been successfully exported to other universities within Spain and South America.

In 2011 the UPV embarked on the EU-subsidised project: 'transLectures' (Silvestre et al., 2012) to develop innovative, cost-effective techniques for the transcription and translation of video lectures. This follows the recommendations of research underlining the importance of transcriptions being available for these video lectures (Fujii, Itou, & Ishikawa, 2006), not only for the purposes of providing subtitles for non-native speakers or people with hearing impairments (Wald, 2006), but also to allow lecture content searches (Repp, Groß, & Meinel, 2008) and other advanced repository management functions. Users might also choose to watch the video lectures with subtitles in order to aid comprehension of a foreign language, even viewing the

transcript and translation simultaneously to help with the language learning process itself. Subtitles can also mean that videos can be followed in both noisy and noise-restricted environments.

As part of transLectures, automatic subtitles in Spanish, English and Catalan have been made available for all videos in the poliMedia repository and are continually being updated as technologies are improved during the course of the project. In order to provide these subtitles, we use a state-of-the-art automatic speech recognition (ASR) toolkit developed as part of the transLectures project, the transLectures toolkit, also known as 'TLK' (The TransLectures-UPV team, 2013), and the well-known machine translation (MT) toolkit, Moses (Koehn, 2007). These toolkits are able to create statistical models and use them to produce high quality transcriptions and translations. Both toolkits were integrated into a complex and distributed system that automatically transcribes the complete poliMedia repository (Silvestre-Cerdà et al. 2013).

Automatic speech recognition (ASR) is a mature research field (Gilbert, Knight and Young, 2008; Jelinek, 1998) which has experienced a major boost over the last two decades. Typically, ASR systems are made up of two statistical models, the acoustic model and the language model (n-gram models). Maximum likelihood and other discriminative techniques can be applied to automatically estimate the parameters of these models using a large corpus of audio and text. The search process provides the most probable transcription hypotheses given the acoustic data. The performance of an ASR system is assessed using Word Error Rate (WER), a common metric used to assess the performance of ASR systems. It measures the difference between the automatic and the manual transcriptions, and is calculated by dividing the minimum number of edits (insertions, deletions and substitutions) required to bring the automatic transcription into line with the manual reference transcription by the total number of words in the reference transcriptions. Our evaluation was carried out on what is called the test set, which includes some four hours of speech with the corresponding reference transcriptions.

As in ASR, current state-of-the-art in MT is dominated by statistical decision systems combining adequate probabilistic models learnt from training data. Given a text sentence to be translated from a source language into a target language, statistical MT (SMT) systems decide its most likely translation by combining language and translation models. Translation models are usually implemented in terms of phrase tables. They are learnt by maximum likelihood estimation, though discriminative training is being increasingly used. System comparison is often carried out on the basis of automatic assessment metrics, such as BLEU or TER for MT systems. BLEU measures the degree of overlap between the automatic translation and a single correct reference translation, comparing isolated words and groups of up to four consecutive words. TER is similar to WER, but also considers any swaps of consecutive word groups required to make system output match the reference translation. A gentle and accessible introduction to current SMT technology can be found in a very recent monograph by P. Koehn (Koehn, 2010).

User evaluations are currently underway in collaboration with UPV lecturers who, having filmed material for the poliMedia repository as part of an earlier *Docència en Xarxa* call, are now trialling the transLectures transcription editing interface.

In this paper we outline the implementation and uptake of poliMedia at the UPV. We also give an overview of the transLectures project, outlining its main goals and motivations, before moving on to a discussion of the three-phase user evaluation stage and the progress made based on user feedback in the area of interface design, functionality and the project's computer-aided interaction component..

poliMedia

poliMedia is a relatively recent service developed and implemented at the Universitat Politècnica de València (UPV) for the creation and publication of educational multimedia content (poliMedia, 2007). Launched in 2007, it is primarily designed to allow UPV lecturers to record pre-prepared mini lectures for use by students to supplement the traditional live lecture. For the most part they consist of concise overviews of a given topic and have a typical duration of around ten minutes. They are also accompanied by time-aligned presentation slides.

So far some 1373 lecturers have recorded 10,174 short video presentations, with levels of participation gaining momentum year on year since 2007: by 2007, 339 lecturers had recorded 1790 videos; in 2008 a further 194 lecturers joined the poliMedia community and 857 videos were recorded; 2009 saw 200 new lecturers sign up and a total of 1145 videos recorded; in 2010, 2011 and 2012, there were 296, 335 and 368 new lecturers, respectively, and some 1204, 1482 and 2129 new videos. So far in 2013, 1449 videos have been recorded and a further 182 lecturers have signed up to the service. These short video presentations are grouped into sets of 4-10 presentations, in order to produce a complete video lecture about a single topic.

This translates into a total running time of 1633 hours of educational material across a range of subjects taught at the UPV, all published online via the poliMedia catalogue. In terms of the uptake by the student body, in 2012 some 30 thousand students accessed these videos (see Fig. 1). Figure 1 also suggests that approximately 15% of UPV students watches video lectures on the poliMedia platform on any given day. We should point out that the videos are viewed at a far higher rate than that at which the transLectures subtitles are activated. This can be explained by the fact that the service only became available a short time ago.

The production process for poliMedia repositories was carefully designed to achieve both a high rate of production and an output quality comparable to that of a television production, but at a lower cost. A poliMedia studio consists of a 4x4 metre room with a white backdrop, video camera, capture station, pocket microphone, lighting and AV equipment including a video mixer and audio noise gate. The hardware cost of this studio stands at around 15,000 euros. We should note that the reduced size of the set means we can obtain a sharper image more easily than if in a standard lecture theatre.

The recording process for poliMedia is quite simple: university lecturers are invited to come to the studio with their presentation and slides. They stand in front of the white backdrop and deliver their lecture, while they and their computer screen (presentation slides) are recorded on two different video streams. The two streams are stacked side-by-side in real-time to generate a raw preview of the poliMedia content, which can be reviewed by the lecturer at any time. These streams are then post-processed; they are cropped, joined (with some overlap) and H.264 encoded to generate an MPEG-4 file, which can be distributed online via a streaming server. All of this is fully-automated and the lecturer can review the post-processed video in a matter of minutes.

The resulting video lectures have a resolution of 1280x720. A number of alternative layouts are available for the user to choose from, as shown in Figure 2.

In addition to the presentation slides, lecturers are requested to provide any metadata and additional textual resources related to the subject of the video lecture. These come into play during the automatic transcription process being developed in transLectures, as part of which the entire poliMedia repository is being regularly re-transcribed using updated versions of the transLectures-UPV open source toolkit, TLK

(The TransLectures-UPV team, 2013). This toolkit is able to create statistical models and use them to produce high quality transcriptions. It also enables some relatively new techniques for adapting the models, which we describe in the following Section.

transLectures

transLectures is the acronym adopted for the EU (FP7-ICT-2011-7) STREP (Specific Targeted Research Project) project 'Transcription and Translation of Video Lectures', in which advanced automatic speech recognition (ASR) and machine translation (MT) techniques are being developed for use in the specific context of large video lecture repositories (Silvestre et al., 2012). The project began in November 2011 and will run for three years. Led by the UPV, the transLectures consortium is made up of three European universities (Institut "Jožef Stefan", RWTH Aachen, Universitat Politècnica de València), three industrial partners (Deluxe Digital Studios, European Media Laboratory GmbH, XEROX S.A.S.) and the Knowledge4All Foundation.

The aim is to produce innovative, cost-effective technologies for the transcription and translation of the vast online collections of video lectures currently emerging within education. Video lectures are being used by universities around the world to enhance, supplement and even revolutionise traditional university curricula (such as MOOCs like Coursera). By adding multilingual subtitles, the content of these videos is opened up to non-native speakers and to the deaf and hard-of-hearing. In addition, raw text data is generated that can be used in combination with other pattern recognition technologies to enable advanced repository management functions, such as lecture search, classification, recommendation and plagiarism detection among others.

The languages or language pairs covered in the transLectures project are English, Spanish and Slovenian for transcription, and bidirectional translation of English<>Spanish and English<>Slovenian, and one way translation of English>French and English>German.

The main scientific and technological goals of transLectures are to prove that the accuracy of these automatic transcriptions and translations can be improved by what we refer to as *massive adaptation* and *intelligent interaction*:

- *Massive adaptation:* This is the process whereby general purpose ASR (and MT) models are adapted to the specific context using lecture-specific variables, such as speaker (Leggetter & Woodland1995) and topic (Martinez-Villaronga, del Agua, Andres-Ferrer and Juan 2013), in order to produce more accurate output. Text data extracted from the metadata and time-aligned presentation slides provided by the lecturers are also used to inform these new 'in-domain' models.

- *Intelligent interaction:* User interaction is not new. At transLectures, however, we adopt an intelligent approach in which the system first identifies which sections of a lecture contain the most errors (Serrano et al. 2013); that is, which sections, based on automatic confidence measures, are most likely to contain errors, and then presents only these sections to the user for correction. These corrections are then fed back into the system and used to re-train the underlying models with a view to avoiding the same errors in the future.

The above ideas are being tested on two real-life case studies: VideoLectures.NET, an online repository with more than 17,000 videos of talks given by top researchers in various academic settings, and poliMedia, described above.

However, these ideas could be applied to any other repositories if additional rich metadata is provided or user supervision is available. In fact, transLectures subtitles have been available on poliMedia, in Spanish and English, since November 2012 and, as part of another side project at the UPV, in Catalan since June 2013. These subtitles will be updated every six months using the latest models and other technologies developed in the lab over the course of the project. For an indication as to student uptake of these subtitles at the UPV, see Figure 1.

Furthermore, user evaluations are currently underway in collaboration with UPV lecturers who, having filmed material for the poliMedia repository as part of an earlier *Docència en Xarxa* call, are now trialling the transLectures transcription post-editing or interaction interface. Specifically, a three-phase evaluation process was set up to explore various modes of interaction, which we discuss in the next Section.

User evaluations

Docència en Xarxa (Teaching Online) is an ongoing incentive-based programme to encourage university lecturers at the UPV to develop digital learning resources based on ICTs. Having already filmed material for the poliMedia repository as part of an earlier call, in 2012/13 lecturers were invited to evaluate the computer-assisted transcription (CAT) tools being developed in transLectures and tested on the poliMedia video repository. Lecturers signing up for this programme committed to supervising[1] the automatic transcriptions of five of their poliMedia videos. These videos were transcribed using TLK, the transLectures-UPV toolkit for automatic speech recognition, which at the time of these evaluations correctly recognised four out of every five words spoken by the lecturer. For evaluation purposes they were allocated across three progressive evaluation stages, described below.

- *First phase:* Lecturers manually supervise the automatic transcription of the first short video presentation from start to end. In this phase, the lecturer plays the short video presentation and the automatic transcription appears, split into synchronised segments of up to 20 words. When they spot a transcription error, the lecturers click on the incorrect segment to enter their correction. The video automatically pauses. In this phase, one video was supervised.

- *Second phase:* In this phase we introduce a word-level computer-aided interaction model. The CAT tool preselects a subset of words within the automatic transcription based on confidence measures (CMs), presenting the lecturer with only those words it considers least likely to be correct. The lecturer supervises these words, playing them in the context of one word before and one word after. Two videos were transcribed in this way.

- *Third phase:* This phase is in fact split into two sub-phases or rounds of evaluation. The first round essentially corresponds to phase two, above, where the lecturer supervises only the sections of the transcript identified as least likely

1 Supervision in this context should be understood as the act of reviewing the automatic transcription, and confirming or correcting the text as necessary; confirming when the suggested text is correct, and correcting when it is incorrect.

to be correct by the CAT tool. The entire video lecture is then automatically re-transcribed on the basis of the lecturer's supervision actions. In a second round, the resulting transcriptions are supervised in full by the lecturer from start to finish, as in phase one. The idea is that these new transcriptions are of a significantly higher quality than the original transcriptions (Sanchez-Cortina, Serrano, Sanchis, & Juan, 2012), so much so that, even counting the time spent in round one, lecturers spend less time supervising the automatic transcriptions. In this phase, the remaining two short video presentations were transcribed.

Feedback was collected after each phase in the form of a brief 10-question satisfaction survey. In addition, the web player (described in more detail below) logged precise user interaction statistics, such as the duration for which the editor window is open, the number of segments (individual subtitles) edited out of the total, the display layout selected; as well as statistics at segment level including the number of mouse clicks and key presses, editing time, and number of times a segment is played. All of this information was used to inform the design of each subsequent evaluation phase and, ultimately, of the player interface itself.

As one of transLectures' main end user or 'prosumer' groups, feedback from university lecturers is fundamental to the outcome of the project. The end goal is a user-friendly platform for post-editing automatic transcriptions that is cost- and time-effective (Luz, Masoodian, & Rogers, 2008).

The CAT tool being tested in this evaluation stage consists of an innovative web player with editing capabilities, complete with alternative display layout options and full keyboard support. It is currently being developed as part of the transLectures project (Valor Miró, Pérez González de Martos, Civera, & Juan, 2012). A screenshot of the player (side-by-side layout) is shown in Figure 3.

First phase: Complete supervision

In the first phase, 20 UPV lecturers supervised the automatic transcription of the first short video presentation in its entirety. The process is straightforward: the lecturers, assigned a username and password, log into the web player to access a private area with their poliMedia videos and select the video they wish to supervise. The web player, shown in Figure 3, is automatically loaded and lecturers can start supervising the transcription straight away.

The web player plays the short video presentation and the corresponding transcription in synchrony, allowing the user to read the transcription while watching the video. When the lecturer spots a transcription error, they press the *Intro* key or click directly on the incorrect segment to pause the video. With the video paused, the lecturer can easily enter their changes in the text box that appears. Lecturers save their work periodically, upon which the transcription is updated and user interaction statistics dumped into a log file.

A preliminary round of phase one was carried out with just two lecturers who volunteered to trial a draft version of the web player. The backgrounds of these two lecturers differed (one from computer science and the other from architecture), which was important in order to avoid opinion biases on issues of interface usability. For instance, the two lecturers presented very different user interaction patterns: the computer science lecturer interacted with the CAT tool primarily using the keyboard, while the architecture lecturer showed a clear preference for the mouse.

Based on the feedback from these first two users, we were able to significantly improve the web player in advance of the launch of phase one proper. Firstly, we shortened the average length of the transcription segments down to 15 words, in line with recommendations from the subtitling industry. This shorter length allows the user to more easily remember what was said in the video and therefore more efficiently correct the words incorrectly recognised by the CAT tool. Secondly, *a search and replace* function was incorporated into the web player, at the suggestion of our computer science lecturer. Finally, both lecturers suggested that correct transcription segments be automatically confirmed once the corresponding video segment has been played, rather than requiring manual confirmation. The remaining 18 lecturers were then asked to supervise the first of their videos using this updated version of the web player.

The most important finding after this first phase was the vast reduction in the time required to produce a transcription for a video lecture. The time spent by lecturers supervising the automatic transcriptions was significantly lower than if transcribing manually from scratch; just over half the time (54%). Indeed, the lecturers' performance became comparable to that of professional transcriptionists (Hazen, 2006), rather than that expected from non-expert transcriptionists (Munteanu, Baecker, & Penn, 2008).

In terms of more qualitative feedback, lecturers valued the simplicity and efficiency of the interface. In the satisfaction survey they collectively scored the web player at 9.1 out of 10 for usability, showing a high acceptance of our CAT prototype as is. That said, one of the lecturers, who had previous professional transcription experience, was unhappy with the interface layout in that it was different to what he was used to working with.

All in all, results were largely positive and, as we hoped, lecturers were able to become familiar with the web player in advance of the next two phases.

Second phase: Computer-aided interaction

In the second phase, we introduced a new interaction protocol called *computer-aided interaction* (Serrano, Giménez, Civera, Sanchis, & Juan, 2013) to find out whether we could further improve supervision times, that is, whether it was possible to make this process even more efficient for the lecturers.

This new interaction mode is based on *confidence measures* (CMs) (Sanchis, Juan, & Vidal, 2012), specifically CMs at the word level. Word-level CMs provide an indicator as to the probable correctness of each word appearing in the automatic transcription. Words with low confidence values are likely to have been incorrectly recognised at the point of ASR and will, therefore, need to be corrected in order to obtain an accurate transcription. With a perfect CM system, the lecturer would only ever supervise (correct) incorrectly-recognised words. In practice they must also supervise (confirm) some correctly-recognised words incorrectly identified as errors. These false positives are unavoidable, since our systems are based on statistical models. They are, however, preferable to false negatives. The idea is that by focusing supervision actions on incorrectly-transcribed words, we can optimise user interaction to get the best possible transcription in exchange for the least amount of effort.

So in this evaluation phase, lecturers were asked to supervise a subset of words preselected by the CAT tool as low confidence, presented in order of probable incorrectness. This subset typically constituted between 10–20% of all words transcribed using the ASR system, though lecturers could modify this range at will to as low as 5% and as high as 40%, depending on the perceived accuracy of the

transcription. Each word was played in the context of one word before and one word after, in order to facilitate its comprehension and resulting correction. Typically, given the starting word error rate (WER) of our automatic transcriptions (10-20%) and an average supervision rate of 15%, our CMs detected around 40% of all real transcription errors.

Figure 4 shows a screenshot of the transcription interface in this phase. In this example, low-confidence words are shown in red and corrected low-confidence words in green. The text box that opens for each low-confidence word can be expanded by the users in either direction in order to modify the surrounding text as required.

For this phase, the computer-aided interaction mode was activated in the web player by default, though lecturers could switch back to the full supervision mode tested in phase one. Analysis of the interaction statistics reveals that only 12 of the 23 lecturers stayed in the computer-aided interaction mode for the supervision of one of their poliMedia videos in full. In the other cases, lecturers switched back to complete the supervision mode. The main reason cited for this was that only by doing so could they be sure to obtain a perfect transcription, something which they professed to value over any time-savings afforded by the computer-aided interaction mode.

In terms of supervision times, the time spent correcting automatic transcriptions in computer-aided interaction mode was reduced to 40% of the time needed for the complete supervision in phase one. However, the resulting transcriptions were not error free, unlike in phase one. That said, the error rate was as low as one in every 10 automatically-recognised words, which is not so far removed from the transcription quality delivered by commercial transcriptions services for academic video lectures (Hazen, 2006).

Feedback from phase two was not as positive as from phase one. Lecturers showed a clear preference for obtaining perfect transcriptions, irrespective of the relative time costs, and insisted that full access be granted to both the video/audio and the transcription. The satisfaction surveys clearly reflected this dislike of the computer-aided interaction mode, preferring a protocol that gives them full control over the end quality of the transcriptions. However, they did seem to embrace the CMs, suggesting that low confidence words be indicated in red font in complete supervision mode also. Overall the system scored 7.2 out of 10 at this stage.

Third phase: Two-round supervision

As indicated, the third phase is divided into two sub-phases or rounds, and is essentially a combination of the previous two phases. First, lecturers supervise a subset of the lowest confidence words, as in phase two, for the remaining two poliMedia videos. The videos are then re-transcribed on the basis of this partial supervision and, in the second round, lecturers supervise the entire re-transcription from start to end, as in phase one. The idea is that the quality of these new transcriptions is higher than that of the original (Sanchez-Cortina et al., 2012), sufficiently so as to allow lower overall supervision times.

In more detail, in the first round lecturers supervised isolated segments of four words in which the last word was the low-confidence word. These segments were presented to the lecturer for supervision in increasing order of confidence (of the last word). The segments kept on being presented to them until one of three conditions were met:

(1) The total supervision time reached double the duration of the video itself; or

(2) No corrections were entered for five consecutive segments; or

(3) Twenty per cent (20%) of all words were supervised.

On average, supervision times during this first round were equal to the duration of the video being supervised, and approximately 15% of incorrectly-recognised words were corrected. These supervision actions were fed into the ASR system used to generate the re-transcriptions, which is the same system used to generate the original transcriptions, but adapted to the lecture- and lecturer-specific variables provided during round one.

Lecturers then embarked on round two, in which they supervised the re-transcriptions from start to end, as in phase one. The time needed to perform this complete supervision was much lower than in phase one because, as expected, the quality of the re-transcription was significantly higher (Sanchez-Cortina et al., 2012) and, consequently, fewer corrections needed to be entered manually: approximately 35% of the incorrectly-recognised words were corrected prior to the start of round two, a combination of the lecturers' efforts in round one and the automatic re-transcription on the basis of their corrections. The end result was a perfect transcription, as in phase one.

In this computer-aided interaction mode, we tried to blend the best outcomes of both previous phases: the perfect end transcriptions of phase one and the shorter supervision times of phase two. Ultimately, this was achieved, though only by a small margin. The complete supervision of the re-transcriptions in round two required 80% of the time needed to do the same task in phase one. However, when supervision times from round one were added, the time-saving with respect to phase one was only very slight (5%).

The main drawback of this model is the two-step process, since lecturers have to put time aside on two separate occasions to supervise the same video. On the whole, a preference (if not requirement) was expressed for the supervision to be carried out in a single session and the corresponding impact on user satisfaction was evident in the average satisfaction survey score for this phase: 7.8 out of 10.

Conclusions

In this paper we have described the poliMedia lecture capture system in place at the Universitat Politècnica de València (UPV) and outlined how the transcription and translation technologies developed as part of the transLectures project are adding value to the poliMedia repository. We paid particular attention to the three-phase internal evaluation stage, in which transLectures tools were deployed in a real-life setting and their usefulness and usability assessed based on feedback from real-life users.

We have reported how the basic user interface trialled in phase one allowed lecturers full control over the quality of the transcriptions of their poliMedia lectures. This mode of interaction scored highly in the satisfaction survey, offering considerable time-savings (54%) relative to transcribing from scratch. We then presented the computer-aided interaction mode tested in phase two which, despite offering even greater time-savings (22% relative to transcribing from scratch), failed to capture the lecturers' interest because it did not lead to perfect end transcriptions. Then, in the third phase, we brought together the best of the previous phases in a two-round supervision process. Here, though, time-savings relative to phase one were minimal and the process less appealing to lecturers, who preferred the simplicity of a single-round supervision process.

However, we should point out that overall transcription accuracy was improved by 35% following an initial supervision of the least confident words, thereby confirming the validity of our use of Confidence Measures (CMs) and computer-aided interaction as a means of improving transcription quality. Ultimately, however, UPV lecturers were not overly interested in this trade off between perfect output and time costs.

Future lines of research will focus on combining the full control allowed in phase one and the use of CMs as in phase two in a way that lecturers find useful and usable. For example, an interface where it was possible to switch seamlessly from complete supervision mode to computer-aided interaction mode, depending on the perceived quality of the automatic transcription, might be better received by lecturers.

Additionally, further user evaluations are planned to test transLectures automatic translation solutions. For these trials, we will need to redesign the user interface to allow side-by-side visualisation of the video, the transcription and the corresponding translation. We will also need to consider different possible interaction protocols, keeping in mind the profile of university lecturer 'prosumers' in so far as their needs and skill sets are not necessarily aligned with those of professional translators, the typical target users of translation technologies.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007–2013) [287755].

References

Fujii, A., Itou, K., & Ishikawa, T. (2006). Lodem: A system for on-demand video lectures. *Speech Communication 48 (5)*, 516–531.

Gilbert, M.,Knight, K., and Young, S (2008). Spoken language technology. *IEEE Signal Processing Magazine*, 25(3), 15-16.

Hazen, T.J. (2006). Automatic alignment and error correction of human generated transcripts for long speech recordings. In *Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006 ICSLP), 1606-1609.*

Jelinek, F. (1998) *Statistical Methods for Speech Recognition*. MIT Press.

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, *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session.*

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Leggetter, C. J., & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech & Language, 9(2), 171-185.

Luz, S., Masoodian, M., & Rogers, B. (2008). Interactive visualisation techniques for dynamic speech transcription, correction and training. In *Proceedings of the 9th ACM SIGCHI New Zealand Chapter's International Conference on Human-Computer Interaction: Design Centered HCI* (pp. 9–16).

Martinez-Villaronga, A., del Agua, M. A., Andrés-Ferrer, J., & Juan, A. (2013, May). Language model adaptation for video lectures transcription. In *IEEE International Conference* on *Acoustics, Speech and Signal Processing (ICASSP)*,, 2013, 8450-8454.

Munteanu, C., Baecker, R., & Penn, G. (2008). Collaborative editing for improved usefulness and usability of transcript enhanced webcasts. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems – CHI 08* (pp. 373–382). New York, NY, USA. doi: 10.1145/1357054.1357117

poliMedia, (2007). The poliMedia tool. URL http://polimedia.blogs.upv.es/?lang=en

Repp, S., Groß, A., & Meinel, C. (2008). Browsing within lecture videos based on the chain index of speech transcription. *Learning Technologies, IEEE Transactions on 1 (3)*, 145–156.

Sanchez-Cortina, I., Serrano, N., Sanchis, A., & Juan, A. (2012). A prototype for interactive speech transcription balancing error and supervision effort. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces* (pp. 325–326).

Sanchis, A., Juan, A., & Vidal, E. (2012). A word-based naïve Bayes classifier for confidence estimation in speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on 20 (2)*, 565–574.

Serrano, N., Giménez, A., Civera, J., Sanchis, A., & Juan, A. (2013). Interactive handwriting recognition with limited user effort. *International Journal on Document Analysis and Recognition (IJDAR)*, 1–13.

Silvestre, J.A., del Agua, M., Garcés, G., Gascó, G., Giménez-Pastor, A., Martínez, A., de Martos, A.P.G., Sánchez, I., Martínez-Santos, N.S., Spencer, R., Miró, J.D.V., Andrés-Ferrer, J., Civera, J., Sanchís, A., & Juan, A. (2012). transLectures. In *Proceedings of IberSPEECH 2012*.

Silvestre-Cerdà, J.A., Pérez, A., Jiménez, M., Turró, C., Juan, A., Civera, J. 2013. A System Architecture to Support Cost-Effective Transcription and Translation of Large Video Lecture Repositories. *Proc. of the IEEE Systems, Man and Cybernetics (SMC), 2013,* 3994-3999.

Soong, S.K.A., Chan, L.K., Cheers, C., & Hu, C. (2006). Impact of video recorded lectures among students. *Who's learning*, 789–793.

The TransLectures-UPV team, 2013. The TransLectures UPV toolkit (TLK). URL http://www.translectures.eu/tlk.

Valor Miró, J.D., Pérez González de Martos, A., Civera, J., & Juan, A., (2012). Integrating a state-of-the-art ASR system into the Opencast Matterhorn platform. *Communications in Computer and Information Science Vol 328 Advances in Speech and Language Technologies for Iberian Languages*, 237–246.

Wald, M., (2006). Creating accessible educational multimedia through editing automatic speech recognition captioning in real time. *Interactive Technology and Smart Education 3 (2)*, 131–141.

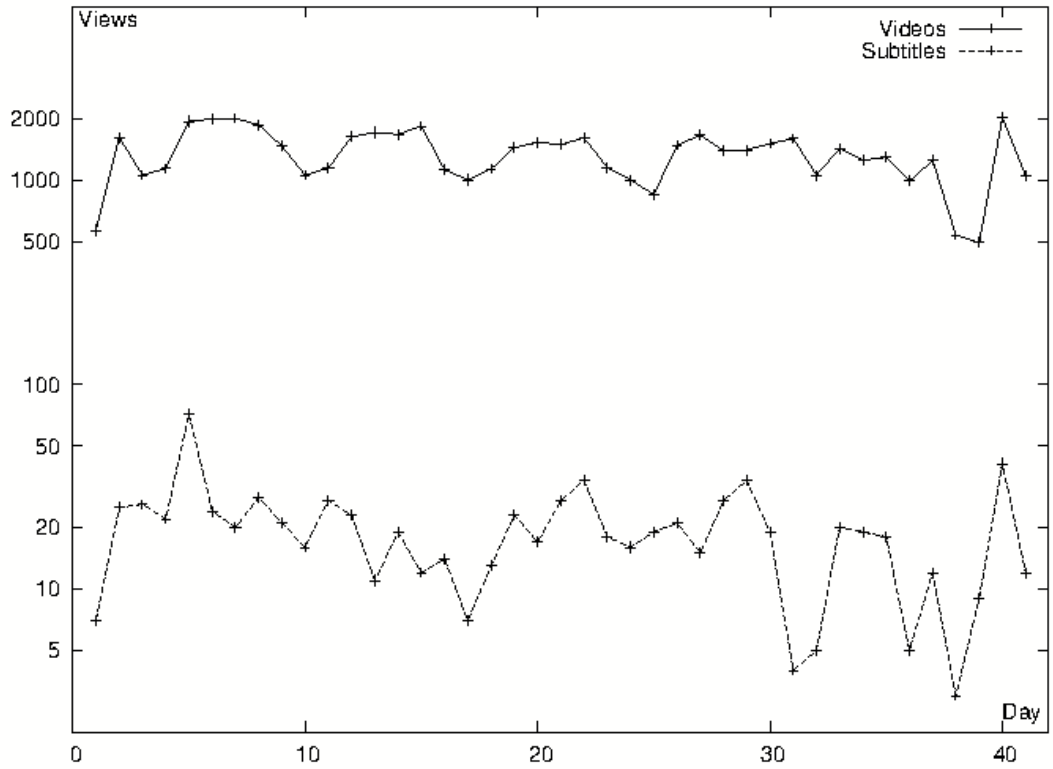

Figure 1. Uptake of poliMedia video lectures and transLectures subtitles in a six-week period at the end of the 2012/13 academic year. As yet both percentage uptake and absolute figures for transLectures subtitles are low, though this is possibly the result of a relative lack of awareness-raising among both teaching staff and students.

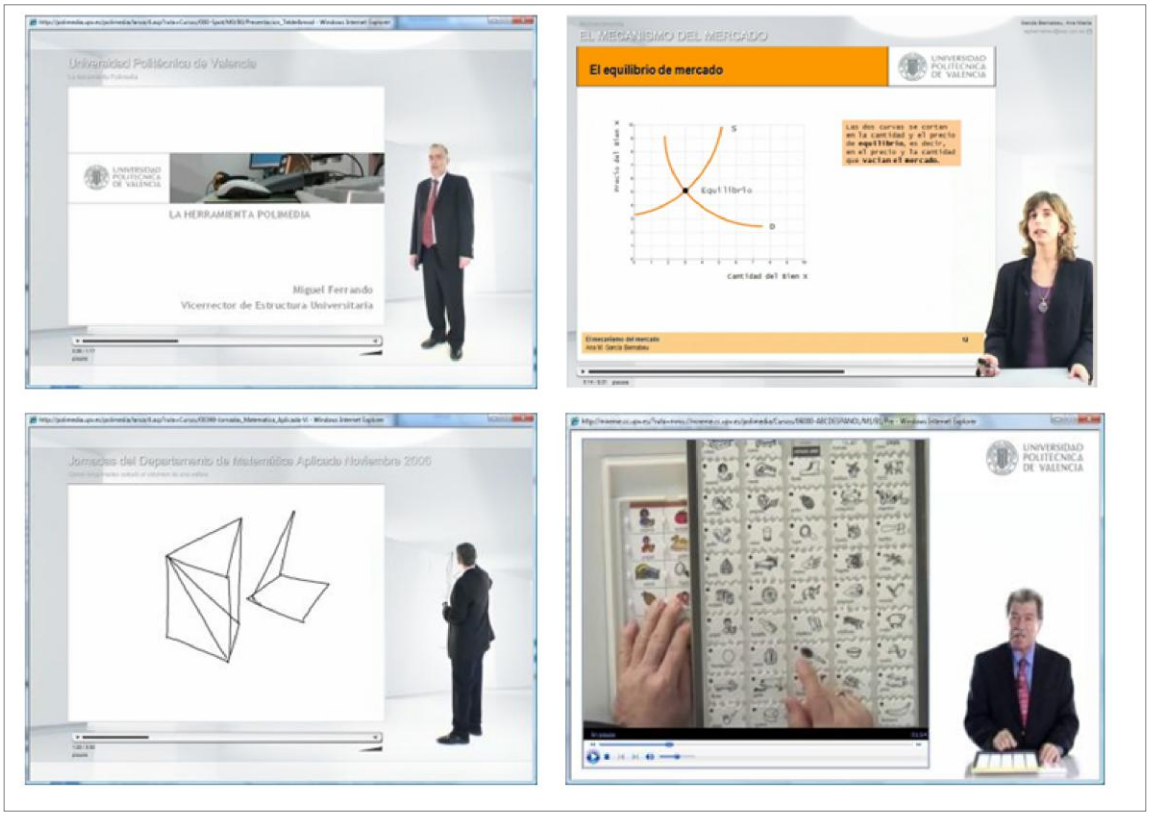

Figure 2. Alternative video lecture layouts for poliMedia. L-r, t-b: full-body shot with presentation slides, half-body shot with slides, whiteboard shot and bird's eye view.

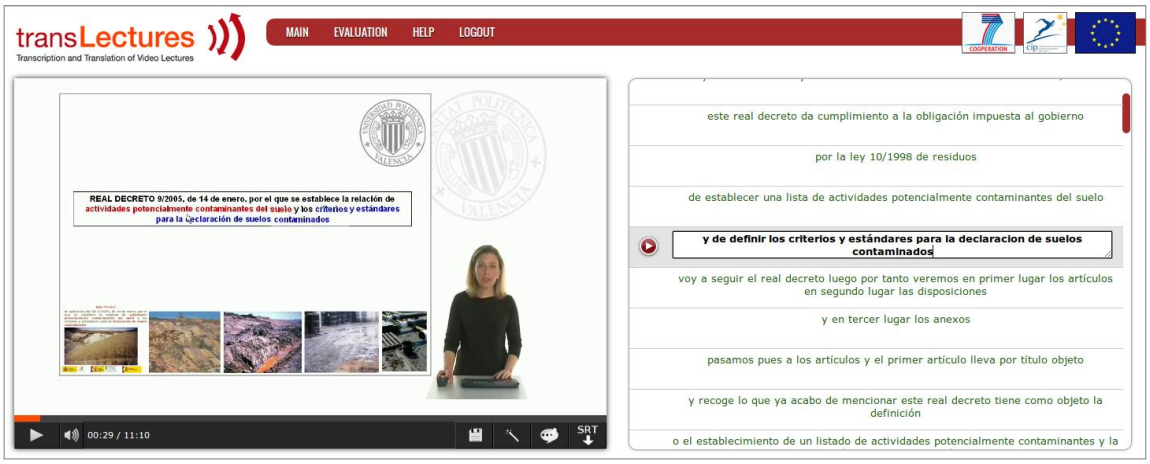

Figure 3. transLectures web player with the side-by-side layout while the lecturer edits one of the segments.

Figure 4. Screenshot of the transcription interface in computer-aided interaction mode. Low-confidence words appear in red and supervised low-confidence words in green. The word being edited in this example is opened for supervision, and the text box can be expanded to the left or right by clicking on << or >>, respectively. Clicking the green check button to the right of the text box confirms the word as correct.

Juan Daniel Valor Miró received his Computer Science degree (2012) from the Universitat Politècnica de València (UPV), Spain, and nowadays is finishing his M.Sc. Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging from the UPV. He is also member of the Pattern Recognition and Human Language Technology research group. His main research interests are in the areas of pattern recognition, speech recognition and user interaction and evaluation.

Rachel N. Spencer is a member of the Universitat Politècnica de València's (UPV) Pattern Recognition and Human Language Technology research group. Having completed her BA Hons in Hispanic Studies at the University of Sheffield in 2006, she moved to Valencia, Spain, and began work as a freelance translator. In 2012 she was contracted by the UPV to work on the European project, transLectures: transcription and translation of video lectures.

Alejandro Pérez González de Martos finished his Computer Science degree in the Universitat Politècnica de València (2012). He continued his studies in Artificial Intelligence and Pattern Recognition and is nowadays finishing his M.Sc. Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging in the UPV. He has been a member of the Pattern Recognition and Human Language Technology research group since 2012, and his research interests include a wide range of pattern recognition applications such as automatic speech recognition, statistical machine translation or digital image analysis.

Gonçal Garcés Díaz-Munío holds a degree in Computer Science and Engineering from Universitat Politècnica de València (UPV) and a BA in Translation and Interpreting from Universitat Jaume I (UJI, Spain). Since 2012, he has been working at the UPV's Pattern Recognition and Human Language Technology research group, as part of the EU FP7 project transLectures: transcription and translation of video lectures.

Dr. Carlos Turró holds a Ph.D. from Universitat Politècnica de València (2003). He is Head of the Media Services department and main developer of the Polimedia service as a system to create Video Learning Objects. Currently the Polimedia system is used from more than ten universities and has more than 9000 learning objects. This system was granted with the Spanish award FICOD 2009 for the promotion of digital services. He has more than 30 publications in international journals and conferences and is working in different EU-funded projects, like the transLectures project, aimed to the automatic and interactive transcription and translation of video lectures and the Rec:All project aiming to research how lectures are currently being captured and used. He is also involved in the Opencast community of Lecture Recordings.

Dr. Jorge Civera is an assistant professor of computer science in the Universitat Politècnica de València (UPV). He received his undergraduate degree from UPV, in 2003 he completed his Master's degree at Georgia Institute of Technology, and in 2008 he received his Ph.D. from the UPV. His Ph.D. thesis was awarded as the best thesis on Computer Science at the UPV 2007-2008. He is co-author of 8 articles in international journals and more than 20 articles in international conferences. He has been involved in several national and international projects, actively participating in the European project TransType2. He is currently leading a Spanish research project on interactive machine translation and speech transcription. He is a member of the Pattern Recognition and Human Language Technology research group and the Spanish AERFAI Society. His research interests include pattern recognition and its application to statistical machine translation, speech recognition, and handwriting recognition.

Dr. Alfons Juan is an associate professor of computer science in the Universitat Politècnica de València (UPV). His research interests include pattern recognition and its application to statistical machine translation, speech recognition, and handwriting recognition. He has led three Spanish projects, the UPV node of the EC Network of Excellence PASCAL2, and is now coordinating the European project transLectures. He is co-author of 14 articles in international journals, more than 60 articles in international conferences and has been the advisor of 6 Ph.D. theses.