transLectures

Transcription and Translation of Video Lectures

D6.3.1: First report on evaluations at the case studies

UPVLC, XEROX, JSI-K4A, RWTH, EML and DDS

Distribution: Public

transLectures
Transcription and Translation of Video Lectures

ICT Project 287755 Deliverable D6.3.1

November 8, 2013

Project funded by the European Community under the Seventh Framework Programme for Research and Technological Development.

Project ref no.	ICT-287755
Project acronym	transLectures
Project full title	Transcription and Translation of Video Lectures
Instrument	STREP
Thematic Priority	ICT-2011.4.2 Language Technologies
Start date / duration	01 November 2011 / 36 Months

Distribution	Public
Contractual date of delivery	October 31, 2013
Actual date of delivery	October 31, 2013
Date of last update	November 8, 2013
Deliverable number	D6.3.1
Deliverable title	First report on evaluations at the case studies
Type	Report
Status & version	v1.0
Number of pages	27
Contributing WP(s)	WP6
WP / Task responsible	UPVLC
Other contributors	
Internal reviewer	Jorge Civera, Alfons Juan
Author(s)	UPVLC, XEROX, JSI-K4A, RWTH, EML and DDS
EC project officer	Susan Fraser

The partners in transLectures are:

Universitat Politècnica de València (UPVLC)
XEROX Research Center Europe (XEROX)
Josef Stefan Institute (JSI) and its third party Knowledge for All Foundation (K4A)
RWTH Aachen University (RWTH)
European Media Laboratory GmbH (EML)
Deluxe Digital Studios Limited (DDS)

For copies of reports, updates on project activities and other transLectures related information, contact:

The transLectures Project Co-ordinator
Alfons Juan, Universitat Politècnica de València
Camí de Vera s/n, 46018 València, Spain
ajuan@dsic.upv.es
Phone +34 699-307-095 - Fax +34 963-877-359

Copies of reports and other material can also be accessed via the project's homepage: `http://www.translectures.eu`

1 Introduction

This is the first report on the evaluations carried out at the project case studies. Commentary in this deliverable will correspond to the internal evaluations deployed under Task 6.3. The external evaluations of Task 6.4 will be covered in the second evaluation report (D6.3.2). As described in the DoW, Task 6.3 is carried out in coordination with Task 5.3 in order to evaluate models, tools and the integration progress in a real-life yet controlled setting.

This document is divided across two main sections describing the evaluations carried out on VideoLectures.NET and poliMedia, respectively. In each case, we describe the entire process, from recruiting users to the conclusions drawn, and reporting the most significant findings at each stage.

2 Internal evaluations at poliMedia

2.1 Introduction

In 2007 the Universitat Politènica de València (UPV) implemented its poliMedia lecture capture system for the cost-effective creation and publication of quality educational video content ([4]). 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.

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. The quality of these subtitles has been internally evaluated with the collaboration of UPV lecturers. Lecturers who filmed material for the poliMedia repository as part of an earlier *Docència en Xarxa* call are now trialling the transLectures transcription editing interface, also known as transLectures player. Specifically, a three-phase evaluation process was set up to explore various modes of interaction.

2.2 poliMedia

poliMedia is primarily designed to allow UPV lecturers to record pre-prepared mini lectures for use by students in supplement to 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.

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. A typical poliMedia video would look something like that shown in Figure 1, though a number of alternative layouts are available for the user to choose from.

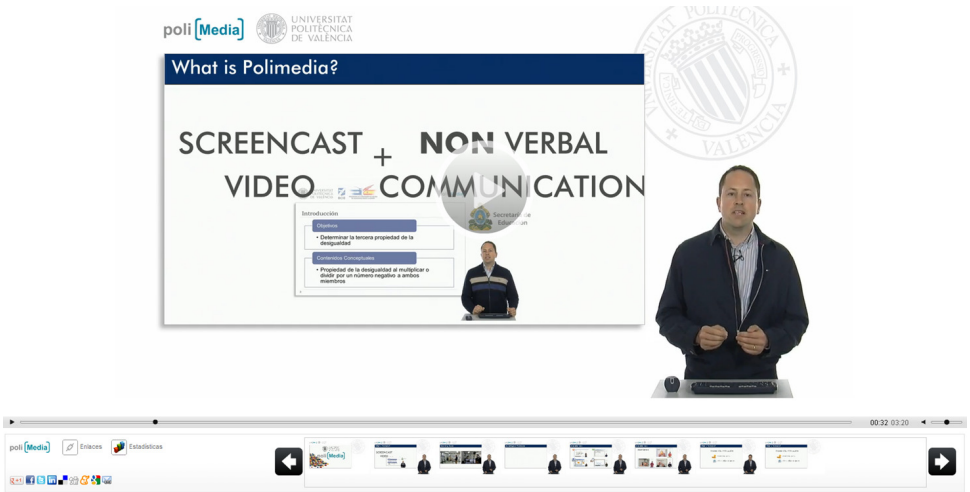

Figure 1: Example of a poliMedia lecture.

2.3 User evaluations

UPV lecturers, having already filmed material for poliMedia, were invited under *Docència en Xarxa* to evaluate the computer-assisted transcription (CAT) tools being developed in

transLectures. 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 [8], which at the time of these evaluations correctly recognised four out of every five words spoken by the lecturer (∼20 WER[2]). For evaluation purposes, lectures were allocated across three progressive evaluation stages, briefly described below.

- First phase: Lecturers manually supervise the automatic transcription of the first video lecture from start to end. In this phase, the lecturer plays the video lecture and the automatic transcription appears, split into synchronised segments of up to 20 words. When they spot a transcription error, they 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 intelligent interaction model. The CAT tool preselects a subset of words within the automatic transcript 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 for each lecturer were transcribed in this way.
- Third phase: This phase is in fact split into two subphases 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 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 the first phase. The idea is that these regenerated transcriptions are of a significantly higher quality than the original transcriptions [5], so much so that, even counting the time spent in round one, lecturers spend less time supervising the automatic transcriptions to get the same result. In this phase, the remaining two videos of each lecturer were transcribed.

Feedback from lecturers was collected after each phase in the form of a brief satisfaction survey. Lecturers were asked to rate various aspects on a scale from 1-10, followed by three open-ended questions where they were free to express their subjective experiences of using the transLectures player (see Appendix A.1). In addition, the web player 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.

As one of transLectures' main end user or "prosumer" groups, feedback from university lecturers is fundamental to the outcome of the project. All of this information was, therefore, used to inform the design of each subsequent evaluation phase and, ultimately, of the player interface itself.

Specifically, the CAT tool being tested and evaluated at this stage by the lecturers consists of an innovative web player with editing capabilities, complete with alternative display layout options and full keyboard support. This player is currently being developed as part of transLectures at the UPV ([9]). A screenshot of the player (side-by-side layout) is shown in Figure 2.

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

[2]Word Error Rate (WER) is the ratio, expressed as a percentage, of the number of basic word editing operations required to convert the automatic transcription into the reference (correct) transcription and the total number of words in the reference transcription.

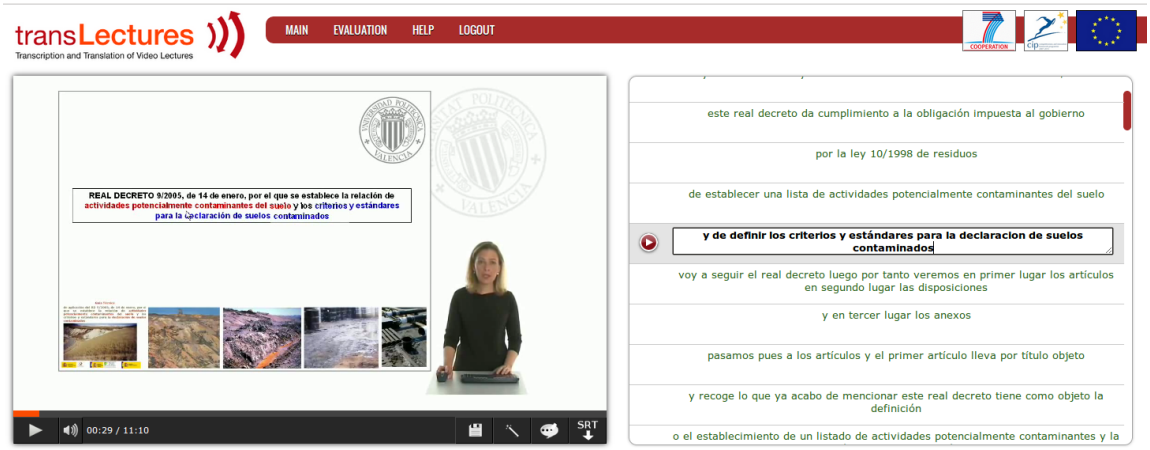

Figure 2: transLectures web player with the side-by-side layout while the lecturer edits one of the segments.

2.3.1 First phase: Complete supervision

In the first phase, 20 UPV lecturers supervised the automatic transcription of the first video lecture 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 2, is automatically loaded and lecturers can start supervising the transcription straight away.

The web player plays the video lecture 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&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.

2.3.2 Second phase: Intelligent interaction

In the second phase, we introduced a new interaction protocol called *intelligent interaction* ([7]) to find out whether we could further improve supervision times, i.e. was it possible to make this process even more efficient for the lecturers?

This new interaction mode is based on *confidence measures* (CMs) ([6]), specifically 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 idenfitied as errors. 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 WER of our automatic transcriptions (10-20%) and an average supervision rate of 15%, our CMs detected approximately 40% of all real transcription errors.

Figure 3 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 in either direction in order to modify the surrounding text as required.

For this phase, the intelligent interaction mode was activated in the web player by default, though lecturers could switch back to the full supervision mode tested in phase one.

2.3.3 Third phase: Two-round supervision

As indicated, the third phase is divided into two subphases or rounds, and is essentially a combination of the previous two phases. First, lecturers supervise a subset of the lowest confidence words, as in the second phase, 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 ([5]), sufficiently so as to allow lower overall supervision times. Figure 4 shows a screenshot of the transLectures web player used in round one.

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.

Figure 3: A screenshot of the transcription interface in intelligent 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.

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 the first phase.

2.4 Evaluation results

In this section we discuss the results obtained in the three phases described above.

2.4.1 First phase: Complete supervision

This first phase involved 20 lecturers, each of whom supervised one of their poliMedia videos as outlined in Section 2.3.1. User behaviour was observed throughout and log files generated by the transLectures player for each transcription session. All of this information was then collated to derive lessons and results.

The preliminary round with two lecturers gave us useful insight into user habits and expectations prior to preparing the first phase for the remaining 18 lecturers. As already reported, the length of the transcription segment was adjusted to make reading, comprehension and subsequent processing easier for the users. A segment length of 10 to 15 words seems most

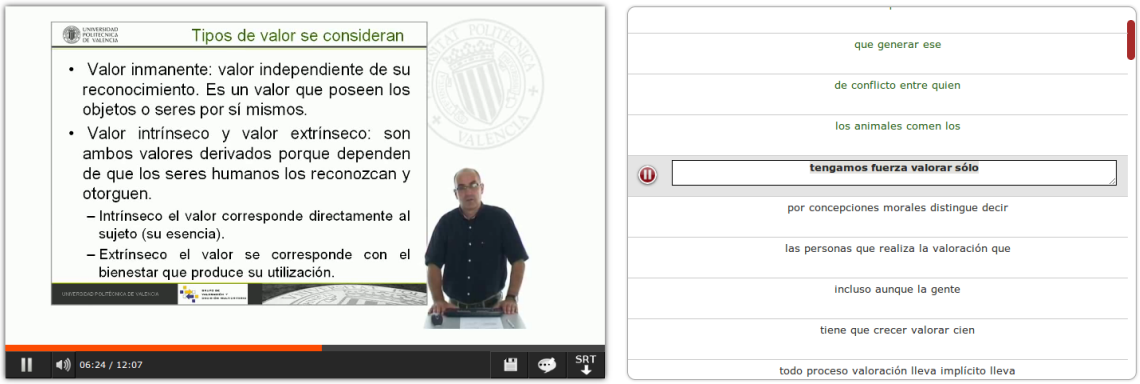

Figure 4: Screenshot of the transLectures web player used in round one of phase three, side-by-side layout. Each segment contains 4 words in which the last word is the low-confidence word.

suitable, though we found in practice that anywhere between 4 and 20 words is also acceptable, in line with the recommendations of the subtitling industry regarding the length of subtitles. With segments of this length, users can more easily remember what was said in the video and therefore amend words incorrectly recognised by the system more efficiently. Secondly, a *Search & Replace* function was incorporated into the web player at the request of one of the lecturers in the preliminary round (see Figure 5). Additionally, both lecturers in this round suggested that transcription segments be automatically validated as soon as the corresponding video segment has been played, making the assumption that the video is only stopped where the transcription contains errors.

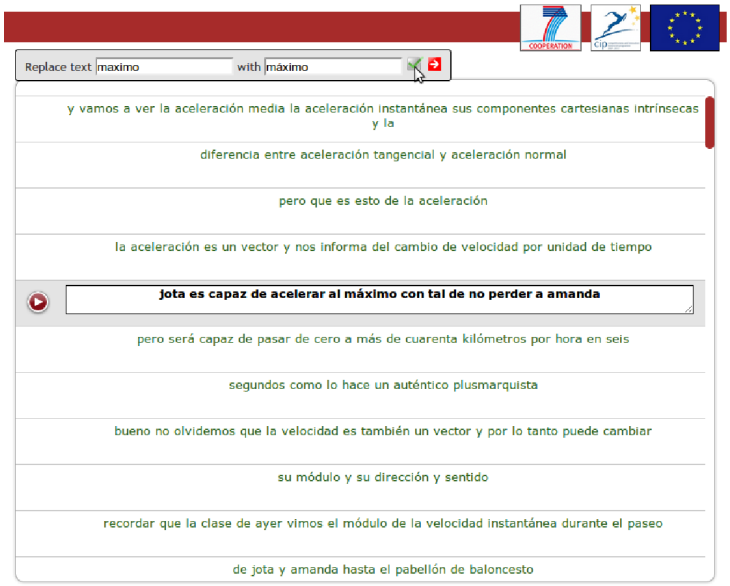

Figure 5: The Search & Replace function incorporated into the web player.

After this preliminary round, the remaining 18 lecturers performed the supervision of their video lectures. Interesting results were recorded in this first phase of internal evaluations. The average real time factor (RTF[3]) was 5.6. If we compare this RTF to that recorded for transcribing from scratch (around 10 RTF for non-expert users [3]), we get a significant decrease of about 50% in user effort to achieve the same transcription quality. Indeed, using the transLectures player, lecturers' performance became comparable to that of professional transcriptionists [1],

[3]In this case, RTF is the ratio between the time devoted to supervising the transcriptions of a video and the duration of said video. So if, for example, a video lasts twenty minutes and its supervision takes, by way of example only, one hour, then the RTF for this video would be 3

rather than that expected from non-expert transcriptionists [2].

The average WER of automatic transcriptions when compared to the lecturer-supervised transcriptions is 16.9 WER, which is consistent with the scientific results reported in deliverable D6.1.2. Figure 6 gives a detailed view at the video level of the first phase, showing RTF as a function of WER. Each cross represents an automatic transcription supervised by a lecturer. As observed, even with automatic transcriptions considered sufficiently accurate as to be usable (i.e. with a WER below 20%), there is no guarantee that non-expert users will spend less than a RTF of 5 on their supervision. This can be compared against performance in Y2QC, where RTF values of below 3 were recorded for the supervision by experts of similar quality automatic transcriptions (see deliverable D6.2.2, Section 3.3)

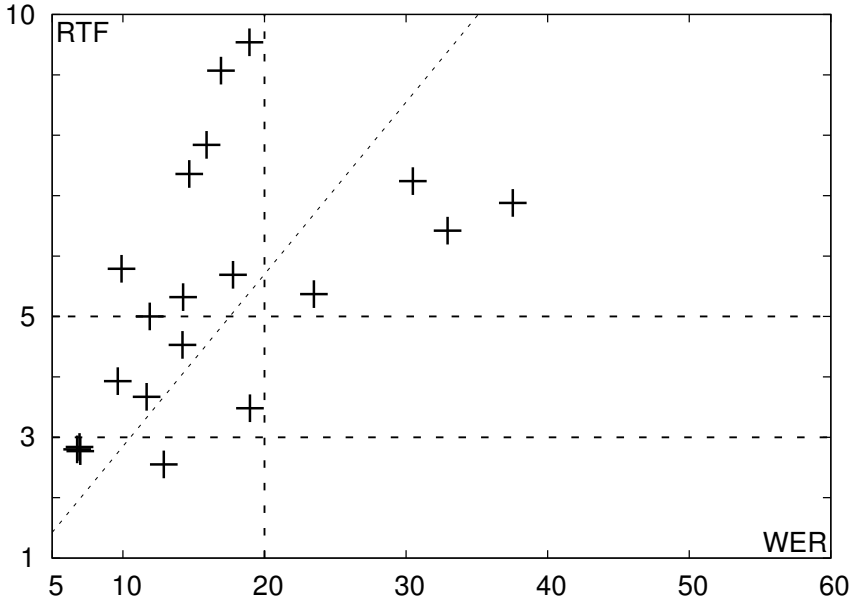

Figure 6: RTF versus WER for transcriptions supervised by lecturers in the first phase.

In order to study how the supervision time devoted by lecturers is distributed across different editing operations (corrections), we fitted a regression model to the data contained in the log files retrieved from the transLectures player after each supervision session. The regression model which best explains our data is defined in terms of the number of words correctly and incorrectly recognised, w_c and w_i, respectively, where t is supervision time:

$$t = a \cdot w_i + b \cdot w_c \tag{1}$$

and a and b are regression coefficients estimated by minimum square error fitting. The model proposed in Eq. 1 closely fit the data with an R-squared value of 0.82, and is statistically significant with a p-value smaller than $2.2 \cdot 10^{-16}$. The value of the coefficient a for incorrectly-recognised words was 4.6 seconds, while the value of b for correctly-recognised words was 1.2. In other words, the average time taken to correct an incorrectly-recognised word is 4.6, while confirming the correctness of a correctly-recognised words takes 1.2, so approximately the duration of the word as spoken. This means that correcting an incorrectly-recognised word takes 4 times longer than confirming a correctly-recognised word. These conclusions are valuable for informing experimental set up in WP4 for the development of a user model that accurately simulates user interaction in order to investigate alternative intelligent interaction modes.

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. Only small improvements

in usability and new features were suggested by lecturers. For example, allowing the user to download the transcription files, change the font size and color, automatically save the transcription file and reduce the initial loading time.

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

2.4.2 Second phase: Intelligent interaction

In this phase lecturers tested the intelligent interaction mode for the first time, but could freely switch back to the complete supervision mode of the first phase at any time.

Analysis of the interaction statistics reveals that only 12 of the 23 lecturers stayed in the intelligent interaction mode for the full supervision of one of their poliMedia videos. In the other cases, lecturers switched back to the complete 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 intelligent interaction mode. As a result the transcription of 18 videos was supervised using the intelligent interaction, while the transcription of 22 videos was supervised in the complete supervision mode.

The RTF of the 22 completely supervised videos (as in the first phase) was 5.2 and WER was 19.5. These figures are in line with those of the first phase. Figure 7 is similar to Figure 6, but the linear adjustment of the points is tighter than in phase 1; that is, lecturers' RTF is more predictable as a function of the transcription WER as they become more proficient transcribers.

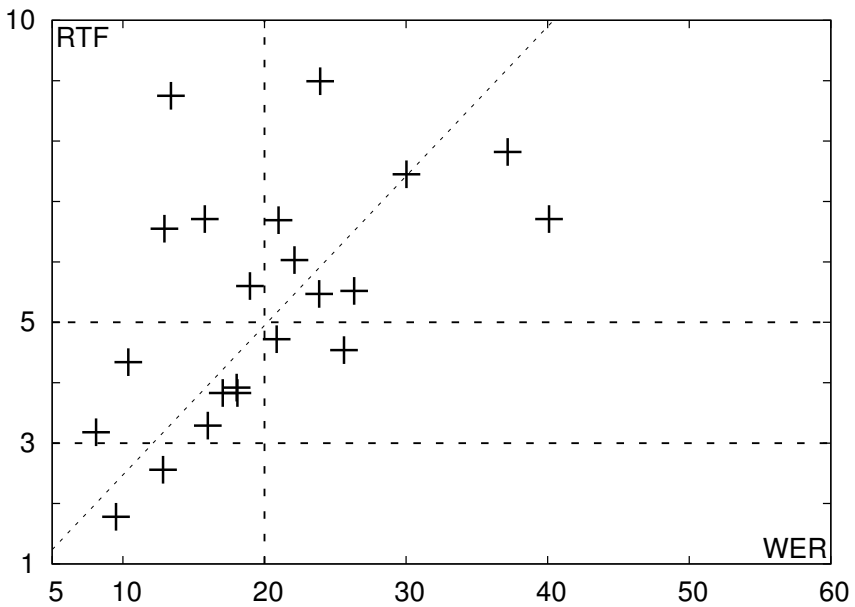

Figure 7: RTF versus WER for transcriptions supervised by lecturers in the second phase.

For those lecturers that opted for the intelligent interation mode, supervision time was reduced to 40% of that needed for the complete supervision in phase one, with an RTF of 2.2. This significant reduction in RTF comes from the fact that lecturers only needed to listen to and supervise those words with the lowest confidence score, and not all of the words in the video lecture transcription.

In Figure 8, and in contrast to Figures 6 and 7, we can see how much WER is reduced in relative terms for a given amount of supervision time, expressed in RTF. For instance, the bottom-right quadrant contains those transcriptions that were significantly improved (relative

WER reduction of over 50%) with a small amount of lecturer effort (less than 2 RTF). Looking at the top-left quadrant, it is unclear whether a greater amount of lecturer effort (above 3 RTF) will translate into greater WER reductions. These conclusions were taken into the third phase, informing the experimental set-up.

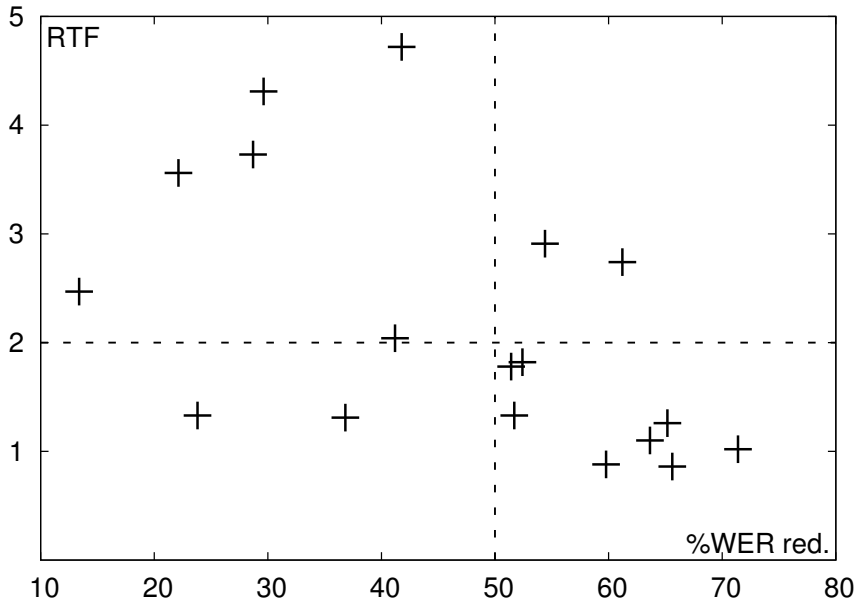

Figure 8: Relation between percentual WER reduction and RTF for each video supervised with intelligent interaction in the second phase.

In this phase the resulting transcriptions were not error free, unlike in phase one. That said, the residual WER of the transcriptions after supervision was as low as 8.0, which is not so far removed from the transcription quality delivered by commercial transcriptions services for academic video lectures ([1]). Overall, WER drops from 14.5 to 8.0 with a supervision time of 2.2 RTF. This indicates that the confidence measures successfullly identify almost half of all incorrectly-recognised words, allowing the intelligent interaction mode to make more efficient use of lecturer supervision effort. However, despite this improvement in supervision time, lecturers clearly stated that they prefer to invest more time supervising the transcriptions in order to obtain error-free subtitles for their videos.

As in the first phase, the time devoted to editing operations was analysed to develop a user model. In this phase, the average time taken to correct an incorrectly-recognised word is 4.9 seconds, 3.2 to supervise a correctly-recognised word. If we compare these results with those extracted in phase one, we observe that the time taken to amend incorrect words is similar. However, the time taken to confirm correctly-recognised words notably increases. This is because, since these words are assumed to be incorrect based on their confidence measures, the interface automatically enters these segments in editing mode, which is takes some time to exit. This extra time cost must be understood as the cost of false positive samples; that is, the system considers a sample to be incorrect when it is actually correct.

Lecturer feedback from the satisfaction surveys was not as positive as in phase one. Lecturers showed a clear preference for obtaining perfect transcriptions, irrespective of the relative time costs afforded by the intelligent interaction mode, and insisted on a protocol that gave 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.

Despite lecturer preferences, the intelligent interaction model based on confidence measures was able to effectively identify incorrectly-recognised words and led to corresponding efficiency

gains in the overall supervision process. We designed the third phase in such a way as to take full advantage of the intelligent interaction mode, while also granting lecturers full control over the final transcription quality.

2.4.3 Third phase: Two-step supervision

The two-step supervision process described in Section 2.3.3 was successfully completed by 15 lecturers on a total of 26 video lectures. It is important to note here that average transcription quality for these videos was measured at 28.4 WER points, so our starting point was worse than in the previous two evaluation phases two phases. This is because the lecturers' first three videos for transcription were selected from among the high-confidence transcriptions, in order to make a positive first impression on the lecturers performing transcription for the first time. In this third phase, lecturers selected the videos for supervision themselves, as was requested during the first two phases. Lecturers naturally wanted to supervise the videos that were currently being watched by students, and not those forming part of old, or even obsolete, courses, and this was not always reflected in the videos allocated based on WER scores.

The average time spent in the first round of this phase (intelligent interation) was 1.4 RTF and approximately 15% of incorrectly-recognised words were corrected, reducing WER from 28.4 to 25.0. As outlined, stricter conditions were defined in this phase than in previous phases regarding the length of time that lecturers should spend on this intelligent supervision stage. The idea was to not overwhelm lecturers with an interaction protocol that they had not really taken to in the previous phase.

After this first round of supervision, the ASR system was adapted via massive adaption for each lecturer, and automatic transcriptions were regenerated, keeping those segments supervised by lecturers. As a result of this adaptation, WER drop significantly to 18.7. This means that after the first round, transcription quality was improved by 34% in exchange for just 1.4 RTF effort on the part of the lecturer.

In the second round, lecturers completely supervised the regenerated transcriptions to obtain perfect end transcriptions as in the first phase. The average RTF devoted to supervision was 3.9, lower than in phase one. Adding the RTF of both rounds together gives a total RTF of 5.3, which is lower than that recorded in the first phase, despite a higher starting WER. A brief summary of the results obtained in this phase is given in Table 1.

Table 1: Summary of results obtained in the two-step supervision phase

Step	WER	RTF
Initial transcriptions	28.4	-
First round of intelligent interaction	25.0	1.4
Massively adapted transcriptions	18.7	-
Second round of complete supervision	0.0	3.9

At Figure 9 can be observed WER versus RTF for the second round of the third phase (similar to Figures 6 and 7). Most transcriptions of this phase were supervised below 5 RTF, and those transcriptions have a WER lower than 20. This plot also reflects the fact that lecturers are learning to supervise faster.

In this phase, 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

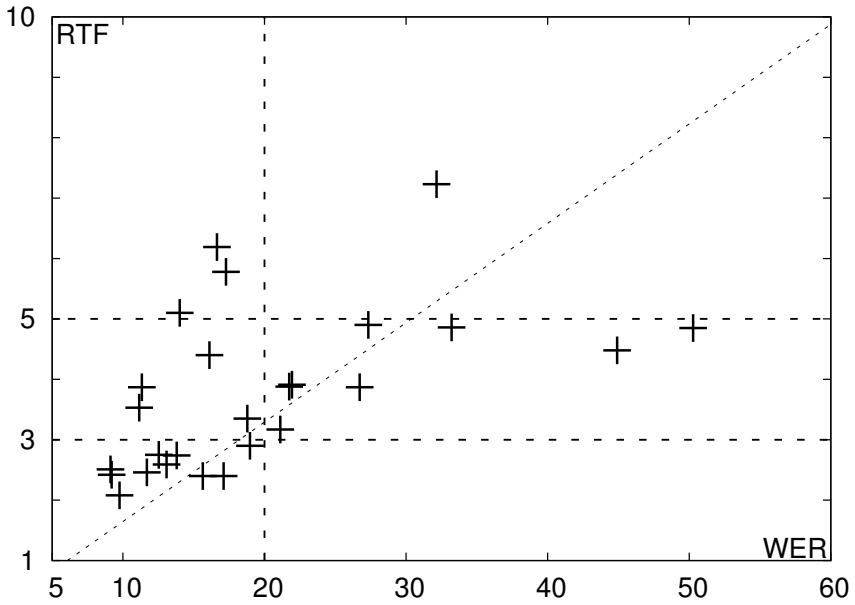

Figure 9: RTF versus WER for the transcriptions supervised by lecturers in the third phase.

supervision times from round one were added, the time-savings with respect to phase one were 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.

2.4.4 Discussion on interaction modes

In the previous sections we have presented the results from the three interaction models used in our internal evalutions at poliMedia. In this section, the three interaction models are compared in terms of WER, RTF and the satisfaction survey (SS) score, in order to get an overview of these results. Table 2 summarizes the results obtained in the three phases.

Table 2: Comparison between all the interaction models evaluated.

Supervision mode	Initial WER	Final WER	Lecturer RTF	SS-score
1st - Manual supervision	16.9	0.0	5.6	9.1
2nd - Intelligent interaction	14.5	8.0	2.2	7.2
3rd - Two-step supervision	28.4	0.0	5.3	7.8

The complete (manual) supervision was the interaction mode most welcomed by lecturers. It is important to note that for most lecturers this was the first time they had supervised automatic transcriptions and, because of the quality of the automatic transcriptions selected, little effort was required on their part. This means their first impression was very positive, as reflected in the satisfaction survey scores.

Meanwhile, the intelligent interaction mode gave good results from the point of view of efficiently reducing WER in exchange for just a small amount of effort in terms of RTF. However, lecturers were really unhappy with not being able to correct the transcriptions to perfection.

Finally, the two-step supervision mode allowed lecturers to obtain perfect transcriptions in exchange for less effort than needed in the complete supervision mode, even when the automatic transcriptions had a higher starting WER. Lecturers saw no clear gain from this two-round process and preferred to stick with the manual supervision.

Figure 10 shows the evolution of RTF as a function of WER across the three phases at poliMedia. As indicated by the linear adjustment to the crosses at each phase, lecturers are gaining experience at supervising transcriptions and improving their RTF figures phase-on-phase.

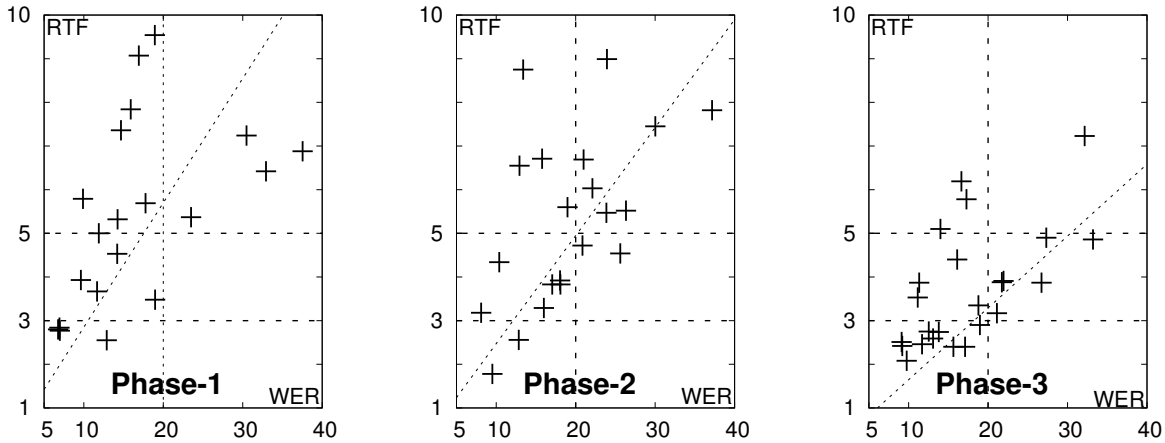

Figure 10: Evolution of RTF as a function of WER across the three phases of internal evaluations at poliMedia.

2.5 Conclusions

In this deliverable we have described the internal evaluations carried out at the poliMedia case study, also analysing the findings. transLectures tools were deployed in a real-life setting, and their usefulness and usability was assessed based on feedback from real-life users (lecturers).

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, and offered considerable time-savings (45%) relative to transcribing from scratch. We then presented the intelligent interaction mode tested in the second phase which, despite offering even greater time-savings (a further 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, time-savings relative to phase one were minimal and the process was 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 34%, thereby confirming the validity of our use of CMs and intelligent interaction as means of efficiently improving transcription quality. Ultimately, however, UPV lecturers were not overly interested in this trade off between perfect output and time costs.

Future work in transLectures will focus on combining the full control allowed in the first phase and the use of CMs as in the second phase 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 intelligent 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 need to redesign the user interface to allow side-by-side visualisation of the video, the transcription and the corresponding translation.

3 Internal evaluations at VideoLectures.NET

3.1 Introduction

In 2002 the Jožef Stefan Institute started with the filming of Solomon seminar[4], which is a series of open public seminars mainly on Artificial Intelligence and general Computer Science topics. From there the VideoLectures.Net website video library was created which at present includes content from 853 events, 11,617 authors and 15,251 lectures - 17,534 videos with a total of 12,220 hours or recorded material. It's a unique resource for high-quality and specialized academic talks and OER[5] resources. VideoLectures.Net is shaping the future of education and as the collection has grown larger, the activities of VideoLectures.Net have become more complex. In 2010 it co-founded the Knowledge 4 All Foundation Ltd, in order to maximize the impact of the online collection and dedicate resources to the creation of tools and intelligent services that complement and enhance its video content.

Knowledge 4 All Foundation Ltd (K4A) is a non-profit organization based in London (UK), formed as a successor to the PASCAL2 Network of Excellence[6], an FP7 research project and the main supporter of VideoLectures.Net. Among the over 60 members' sites of the Foundation are some of the most important research and development centers in the field of artificial intelligence in Europe. The aim of the Foundation is to reduce the current gap between new trends in education on the one hand, and advanced technologies in artificial intelligence on the other, with the ultimate goal of securing the future of open education. It has also developed an innovative series of video journals, and supports the development of advanced technologies for VideoLectures.Net. K4A works closely with the OpenCast Community and OpenCourseWare Consortium[7], both of which operate in the field of open source video capture systems and open educational resources, and provides intelligent solutions for both.

As part of transLectures, automatic subtitles in Slovene and English have been made available for 10000 videos in the VideoLectures.Net library (9242 English and 767 Slovenian) and are continually being updated as technologies are improved during the course of the project. The quality of these subtitles has been internally evaluated with the collaboration of undergraduate and graduate students of the Department of Translation studies, Faculty of Arts, University of Ljubljana. The same students were responsible for a two-phase evaluation process, setup at Jožef Stefan Institute with the help of UPV and monitored by Knowledge 4 All Foundation, to explore distinctive modes of interaction.

3.2 VideoLectures.NET

VideoLectures.Net was designed to allow lecturers around the globe to improve their profiles by being filmed and their talks be placed online. It also functions as a gateway for a mass of structured, peer reviewed and research oriented conferences, workshops and seminars mainly in computer science but also represented by all other sciences, including arts and humanities. The principle is to record lectures in real, non-controlled environments for use by scholars in supplement to the traditional conference lecture, flipped classroom in a blended learning approach. Mainly they consist of overviews of a specific, often published topic and have average cross-repository duration of around 35 minutes. Like in poliMedia, they are also accompanied by time-aligned presentation slides. They feature a download box, comment box, description, detailed multi-subject categorization, recording and publishing date, number of views, author metadata (name, surname, institution) and recommendation system called LaVie created within the transLectures dissemination package WP7 (see Figure 11).

[4]Solomon seminar video talks `http://videolectures.net/solomon/`
[5]Open Educational resources
[6]PASCAL2 website `http://www.pascal-network.org`
[7]OCWC consortium `http://www.ocwconsortium.org`

Figure 11: Screenshot of a VideoLectures.NET lecture.

3.3 User evaluations

Contrary to poliMedia there is only a handful of lecturers at VideoLectures.Net that are in close contact to the repository, which basically functions as a filming service for lecturers, event organizers and other content providers. Lecturers are rarely contacted after filming, only for a check-up and preview, as this typically presents an overhead for them, many being even affiliated to different institutions. We decided therefore to invite the translation undergraduate and graduate students on the account of their agility, enthusiasm and early expertise building to evaluate the computer-assisted transcription (CAT) tools being developed in transLectures project. The videos chosen for evaluation were transcribed by EML using their internal models for English and Slovenian automatic speech recognition. Students were allocated across two evaluation phases. Here is a brief description of the two phases:

- First phase: 5 students manually supervised the automatic transcription of 21 Slovene and 26 English video lectures each from start to end. In this phase, complete supervision was performed and log files were generated from user interaction. The students played the video lectures with the transLectures player, enabling them to view the automatic transcription. When they spotted a transcription error, they clicked on the incorrect segment and entered their correction. The video automatically paused and was resumed after a successful edit.
- Second phase: this phase was similar to the second round of the third poliMedia phase from the user interaction point of view. Same 5 students supervised the automatic transcription of 7 Slovene and 37 English video lectures from the same lecturers from the first phase, but different lectures. We assisted students by marking/highlighting the words with the lowest confidence measure with a different color (red). Otherwise the editing process was the same as in the first phase.

Much time was spent on preparing the testing environments and synchronizing between the UPV, JSI and K4A teams on software bugs, fixes, servers' settings and other administrative preparations with the Faculty of Arts. Due to these factors, the testing began later then at poliMedia. The students were given specific transcription guidelines, designed and prepared for Task 6.3. Same as at poliMedia, feedback from students was collected after each phase in the form of a brief 10-question satisfaction survey and free text comments. The web player logged precise user interaction statistics.

3.3.1 First phase: Complete supervision

In the first phase, 5 students supervised the automatic transcription of 10 video lectures each, totaling 47 videos (21 Slovene lectures and 26 English lectures). Same as at poliMedia, the students were assigned a username and password, logged into the web player to access a private area with the selected VideLectures.Net videos, where they opened the videos they were assigned to supervise in advance (see Figure 12). The web player, shown in Figure 13, was automatically loaded and the supervision of the transcriptions started straight away.

The students were summoned for a live meeting at the Faculty of Arts where the project was described to them, followed by an online meeting where they were advised to try the evaluation on one video and report any problems. In line with the first phase at poliMedia, the students showed a clear preference for using the mouse instead of the keyboard.

Based on the feedback from the first trial we shortened the average length of the transcription segments from longer segments of 50 words down to 10-15 words. This shorter length allowed the students to visually follow what was said and react faster in order to correct the incorrectly recognized words by the CAT tool.

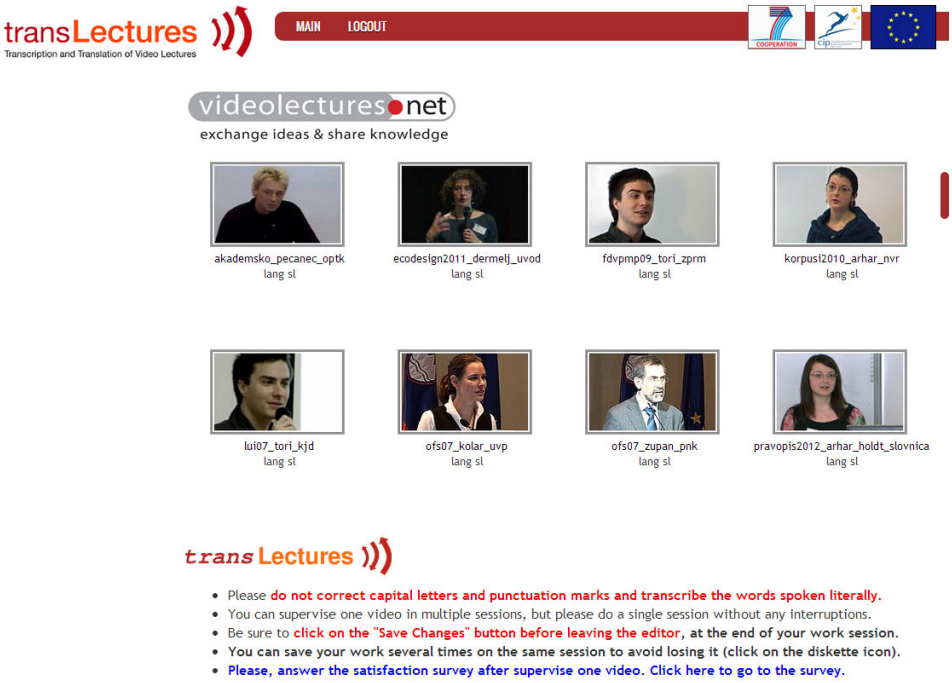

Figure 12: Snapshot of first screen after logging in showing videos to be supervised.

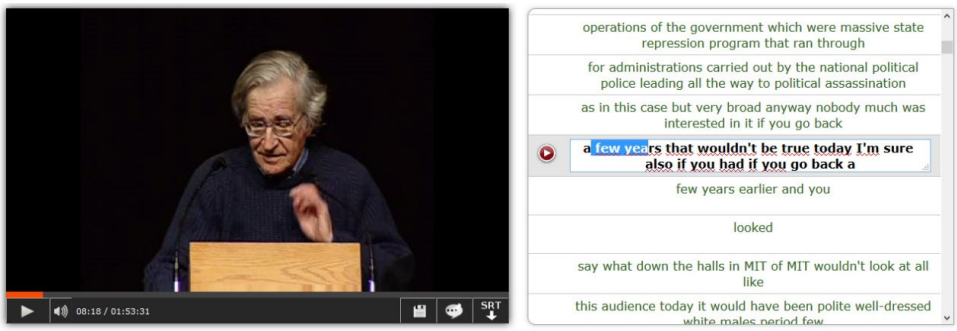

Figure 13: Screenshot of transLectures player while supervising English automatic transcriptions for the first phase of internal evaluations at VideoLectures.NET.

3.3.2 Second phase: Assisted complete supervision

Based on the feedback from the poliMedia second phase results with the lecturers' engagement in the intelligent interaction protocol, it was decided that this intelligent interaction phase should be skipped at VideoLectures.Net. Instead we decided we will only mark the words with the lowest confidence measure with a different color (red), so the students would be able to easily spot the possibly wrong words.

For the second phase the same students' group was used. Lectures for this round were selected from the same lecturers as in the first phase, but with different videos. Similarly as at poliMedia the videos were re-transcribed with an improved ASR system based on the complete supervision from the first phase. The videos from the same lecturers were re-transcribed so that massive adaptation techniques could be put into practice for this second phase.

In this second phase, students supervised the entire re-transcription from start to end, as in phase one. The idea was that the quality of these new transcriptions would be higher and supervision time accordingly shorter. Figure 14 show the evaluation environment and the web player in second phase.

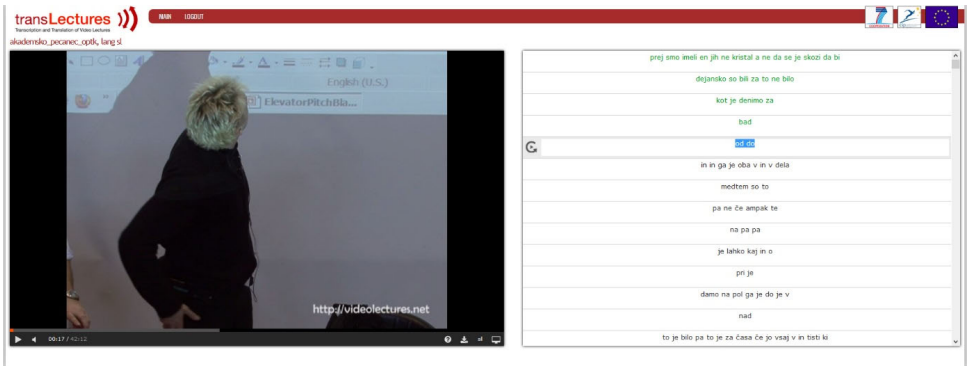

Figure 14: Screenshot of transLectures player while supervising Slovene automatic transcriptions for the second phase of internal evaluations at VideoLectures.NET.

3.4 Evaluation results

This section discusses the evaluation results obtained in the two phases previously described.

3.4.1 First phase: Complete supervision

This phase involved 5 students, who supervised 47 videos as explained in Section 3.3.1. Equally to poliMedia the annotations on direct observation of users' behaviour and logs generated by the player for each transcription session were processed to derive lessons and results.

The first round of preliminary testing of the platform provided feedback on the length of the transcription segments which was adjusted to 10 to 15 words, and an addition of the very valuable Search & Replace function was incorporated based on the poliMedia experience. After this preliminary round, the evaluation began.

Table 3 shows overall RTF and WER values for the supervised English and Slovene transcriptions. If we compare those figures to the performance of experts in Y2QC, where RTF values of approximately 5 for both languages, we can observe that student performance was far from that of professional transcriptionists.

Language	WER	RTF
English	39.2	12.3
Slovene	41.3	12.4

Table 3: Overall WER and RTF figures for English and Slovene transcriptions in the first phase.

Figure 15 provides a detailed view at the video level for English and Slovenian transcriptions in the first phase, showing RTF as a function of WER. Each cross represents an automatic transcription supervised by a student. Most transcriptions offered WER above 30 which significantly increase the time needed to supervise them. However, most transcriptions with WER below 30 guaranteed that supervision times were less than 10 RTF.

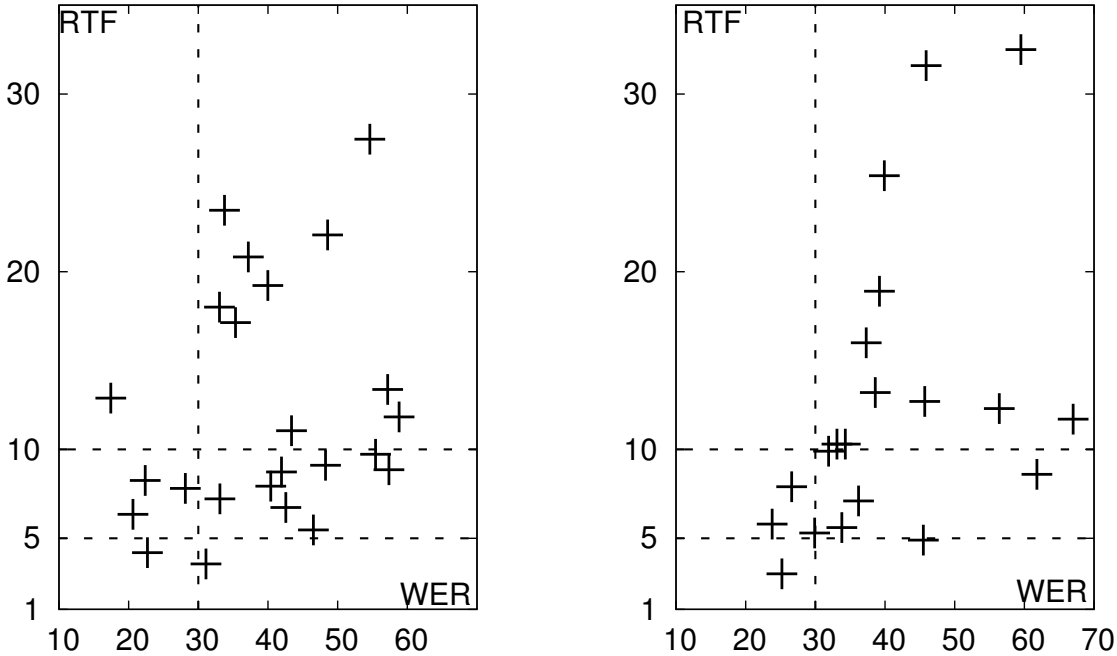

Figure 15: Evolution of RTF as a function of WER for English (left) and Slovenian (right) transcriptions in the first phase.

In the satisfaction survey in terms of accessing the player, the students had some comments about the player (see Appendix A.3) and scored the web player at 6.2 out of 10 for usability. Allowing the user to download the transcription files, change the font size and color and save the transcription file, proved to be of some help as was in poliMedia's case.

3.4.2 Second phase: Assisted complete supervision

After the first round of supervisions, students completely supervised the videos from the same lecturers, but transcribed with an improved ASR system. We expected that the quality of these new transcriptions would be higher and supervision time accordingly shorter.

Table 4 shows overall RTF and WER values for the supervised English and Slovene transcriptions. Although, RTF values are similar to those of the first phase, WER figures are significantly lower than in the first phase after applying massive adaptation techniques to the Slovene and English ASR systems.

Language	WER	RTF
English	26.4	9.9
Slovene	38.5	8.1

Table 4: Overall WER and RTF figures for English and Slovene transcriptions in the second phase.

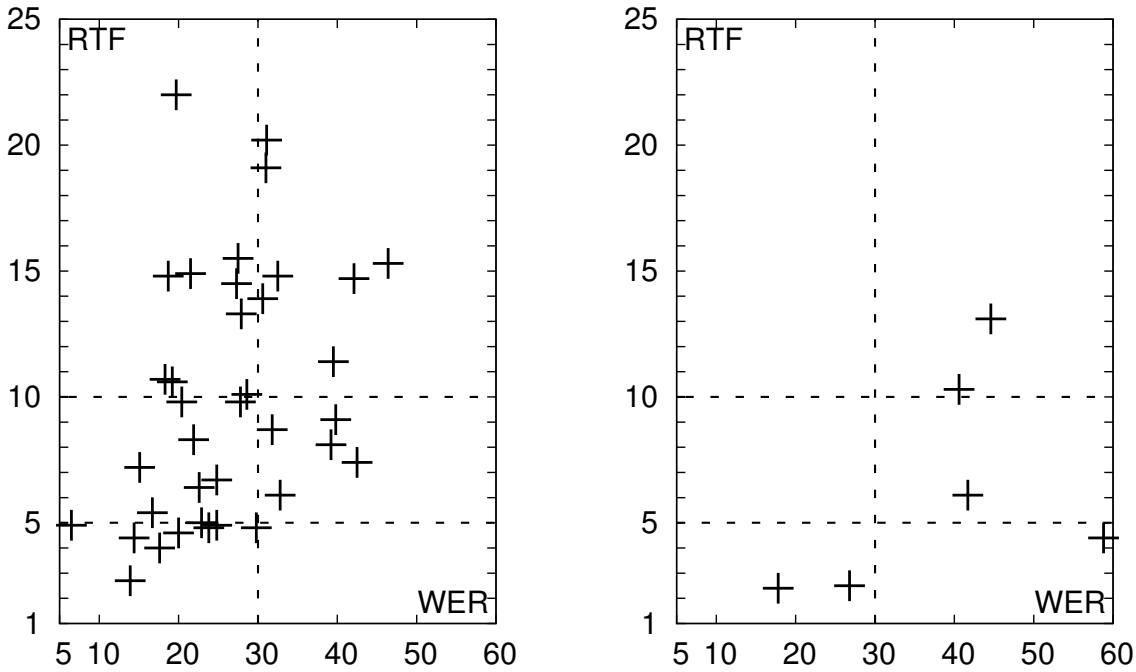

Figure 16: Evolution of RTF as a function of WER for English (left) and Slovenian (right) transcriptions in the second phase.

Figure 16 gives a detailed view at the video level for English and Slovene in the second phase. Slovene transcriptions with WER below 30 were supervised in less than 3 RTF, but this was not the case for English transcriptions.

As derived from the results, the idea of marking with red low confidence words was not helpful for reducing the time needed for supervision (RTF). This observation was also confirmed by students (see Appendix A.3). However, after attending students' comments in the transLectures player, their satisfaction was higher than in the previous phase scoring the web player at 7.4 out of 10 for usability.

It is important to note that most of the students were already acquainted with the subtitling industry and with subtitles in general. Nevertheless, WER values above 20 caused that evaluating and editing automatic transcriptions required considerable effort from their side.

3.5 Conclusions

In case of VideoLectures.Net we decided together with UPV to skip the phase with the word-level intelligent interaction mode. We used the same transLectures player in real-life setting, but our videos were much longer (approx. 3 times). We also supervised automatic transcriptions of two languages (English and Slovene) instead of just one.

The selected interactive model was well accepted by students, although because of longer length of videos they noticed some slowdown in operations. This had a negative impact on the satisfaction level.

The marking of words with lower confidence with a different color proved not to be so useful as the quality of the automatic transcriptions was not so good overall.

Students also suggested some improvements to the web player/editor, which could speed up the process of supervising (see Appendix A.3).

References

[1] Timothy J. Hazen. 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)*, 2006.

[2] Cosmin Munteanu, Ron Baecker, and Gerald Penn. Collaborative editing for improved usefulness and usability of transcript-enhanced webcasts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 373–382, New York, NY, USA, 2008. ACM.

[3] Cosmin Munteanu, Gerald Penn, and Xiaodan Zhu. Improving automatic speech recognition for lectures through transformation-based rules learned from minimal data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 764–772. Association for Computational Linguistics, 2009.

[4] poliMedia. The polimedia tool, 2007.

[5] Isaias Sanchez-Cortina, Nicolás Serrano, Alberto Sanchis, and Alfons Juan. A prototype for interactive speech transcription balancing error and supervision effort. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 325–326. ACM, 2012.

[6] Alberto Sanchis, Alfons Juan, and Enrique Vidal. A word-based naïve bayes classifier for confidence estimation in speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2):565–574, 2012.

[7] Nicolás Serrano, Adrià Giménez, Jorge Civera, Alberto Sanchis, and Alfons Juan. Interactive handwriting recognition with limited user effort. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–13, 2013.

[8] The TransLectures-UPV team. The TransLectures UPV toolkit (TLK). `http://www.translectures.eu/tlk`, 2013.

[9] Juan Daniel Valor Miró, Alejandro Pérez González de Martos, Jorge Civera, and Alfons Juan. Integrating a state-of-the-art asr system into the opencast matterhorn platform. In *Advances in Speech and Language Technologies for Iberian Languages*, volume 328 of *Communications in Computer and Information Science*, pages 237–246. Springer Berlin Heidelberg, 2012.

A Detailed results from satisfaction surveys

A.1 Questionnaire

The first 10 questions are listed below:

- I am satisfied with how easy it is to use this system.
- I can effectively complete my work using this system.
- I can complete my work quicker than doing it from scratch.
- I feel comfortable using this system.
- It was easy to learn to use this system.
- The help information of this system is clear.
- The organization of information on screen is clear.
- I like using the interface of this system.
- This system has all the functions that I expect to have.
- Overall, I am satisfied with this system.

The last 3 open questions are those listed below:

- If you were to add new features to the player, which ones would be?
- If you had to work daily with this player, what would you like to change?
- Any additional comment.

Questionnaires used at both case studies were the same, but VideoLectures.NET added an additional question for Phase 2:

- Was the marking of wrong words with the red color helpful? (Yes/No)

A.2 Results of the satisfaction survey at poliMedia

Table 5 summarises the score of each question from the questionnaire detailed above in each phase. Average values appear in the last row labeled as *Avg*. As observe the best valorated system by lecturers was that of the first phase. The system used in the second phase receive lower overall scores in general, being the score of the ninth question "This system has all the functions that I expect to have" significantly low. In the third phase, even though the supervision time involving the intelligent interaction mode was minimised, the scores were also lower than in the first phase, but a bit higher than in the second phase.

It is important not to undervalue the novelty effect of the transLectures system in the first phase, since most lecturers had not been exposed to such kind of system before and their first impression was greatly positive. However, lecturers in the second and third phase could not help comparing these systems to that of the first phase, so they became more critical as reflected by figures in Table 5.

Regarding the answers to the three open questions, the feedback in the first phase attained small usability details and new features such as:

Table 5: Detailed results of the satisfaction survey for each phase

Question	Phase 1	Phase 2	Phase 3
1	9.4	7.8	7.5
2	9.4	6.7	7.7
3	9.2	6.6	7.4
4	9.0	6.5	7.3
5	9.7	8.1	8.6
6	8.7	8.1	8.5
7	9.0	8.4	8.7
8	9.0	6.9	7.4
9	8.6	5.6	7.1
10	9.0	6.9	7.4
Avg	9.1	7.2	7.8

- Limit segments to 20 words.
- Allow to change the font size and color.
- Allow to download transcriptions.
- Auto-check segments once played.
- Search and replace functionality.
- Automatically save transcriptions.
- Reduce the initial loading time.

In the second phase, lecturers' feedback clearly pointed out that the intelligent interaction model required an in-depth review:

- Allow to edit words out of the intelligent interaction guide.
- Unlimited use of the expansion arrows to allow correct entire segmentd.
- Automatically validate all the occurrences of a correct word in the transcription.
- Allow to go back through intelligent interaction changes.

Finally, in the third phase lecturers' suggestions indicated that, although the two-step supervision provides a little improvement in user effort and transcription quality, this improvement does not fully justify the supervision in two rounds instead of one. Representative comments from lecturers were:

- I spent as much time supervising the last two videos (third pase) as devoted to the first video (first phase).
- Transcription quality has been improved compared with that of the first phase.

In conclusion, the use of satisfaction surveys over the three phases has been a valuable tool to capture lecturers' feedback that reverted in the improvement and refinement of the transLectures system.

A.3 Results of the Satisfaction Survey at VideoLectures.NET

Table 6 summarises the score for each question from the questionnaire for both phases. Average values appear in the last row labeled as *Avg*.

Question	Phase 1	Phase 2
1	7.5	6.3
2	5.5	6.3
3	4.5	6.6
4	7.5	6.6
5	7.0	9.0
6	7.5	9.0
7	7.0	9.3
8	7.0	6.3
9	2.5	8.0
10	5.5	6.3
Avg	6.2	7.4

Table 6: Detailed results of the satisfacton survey for each phase

As in poliMedia, the evaluation system presented to students in the first phase was new to them, so some adapting to the user interface was needed. That is reflected in lower satisfaction score for the first phase. In the second phase the web player was improved, so the satisfaction score is higher,

Web player/editor also showed signs of slowdown during the longer editing sessions/video playback. This was not noticeable in poliMedia case, which videos are in average up to 3 times shorter.

Some suggestions and feature requests from the supervisors:

- Faster loading of video
- Better navigation through the text for easier editing (more keyboard functions)
- Larger display for text editing
- Option to hide the video: bigger text editing part, listening only audio
- Ability to work/edit offline

These suggestions will help us improve the user experience with the transLectures system.