transLectures

Transcription and Translation of Video Lectures

D6.2.2: Second report on quality control

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

Distribution: Public

transLectures
Transcription and Translation of Video Lectures

ICT Project 287755 Deliverable D6.2.2

November 15, 2013

SEVENTH FRAMEWORK PROGRAMME

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

Information Society Technologies

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 15, 2013
Deliverable number	D6.2.2
Deliverable title	Second report on quality control
Type	Report
Status & version	v1.0
Number of pages	37
Contributing WP(s)	WP6
WP / Task responsible	DDS
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

Executive Summary

This deliverable reports on the second quality control process applied on automatic transcriptions and translations in VideoLectures.NET and poliMedia.

Contents

1 Introduction

As part of WP6, a small but significant amount of transcribed and translated data after major upgrades of project models and tools was scheduled for manual supervision by project experts by DDS (for VideoLectures.NET) and UPVLC (for poliMedia). This is the second report on the quality of said transcriptions and translations.

The next section describes which tool, guidelines and lectures were selected for the manual supervision of VideoLectures.NET and poliMedia in this second quality control. Sections 3 and 4 are devoted to the evaluation of transcription and translation quality, respectively. Conclusions are drawn in Section 5. Further details about the transcription and translation guidelines followed by experts during supervision are described in Appendices A.1, A.2 and A.3. Finally, the complete sets of selected lectures are detailed in Appendix A.4.

2 Selection of lectures and assignment of tasks

The first step towards this deliverable was to select a subset of automatic transcriptions and translations from the complete set of transcriptions for Videolectures.NET and poliMedia. The quality control process would be performed by DDS for Videolectures.NET and by UPVLC for poliMedia. The new supervised data are also to be added to the current transcriptions and translations in WP2.

Discussions on the selection of the transcribed and translated lectures for manual evaluation began among relevant project partners in July 2013. After thorough discussion, they agreed on what follows.

2.1 Tool for transcription and translation evaluations

It was agreed that the tool to be used for the evaluation of the automatic transcriptions and translations would be the transLectures player, for the following reasons:

- Internally developed tool that can be easily adapted to transLectures needs and smoothly integrated with the transLectures web service.
- Support to allow side-by-side visualisation of the video, the transcription and the corresponding translation.
- User interaction and timing statistics are automatically collected for each session..
- Alternative editor layouts and comfortable selection of source/target languages.
- transLectures TTML format fully supported.

Figure 1 shows the transLectures player while supervising an English transcription. Video and transcriptor editor are displayed side-by-side in synchrony. Supervised transcriptions are denoted in green. Figure 2 illustrates the transLectures player in translation mode for English into Spanish. Video, source and target texts are shown side-by-side in synchrony. Display layout could be easily modified by the user so that source and target texts are located side-by-side below the video being displayed.

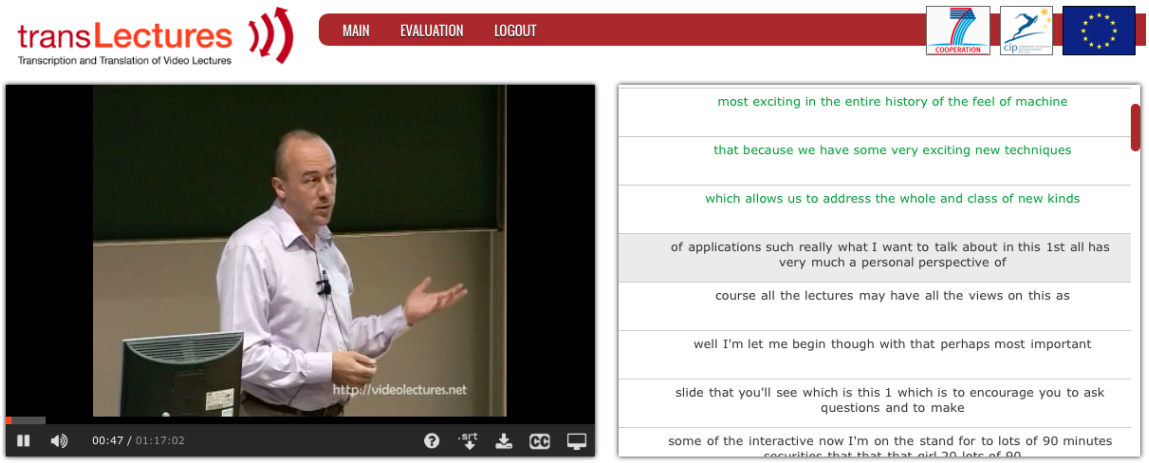

Figure 1: transLectures web player in transcription mode for English with the side-by-side layout when the expert has already supervised three segments.

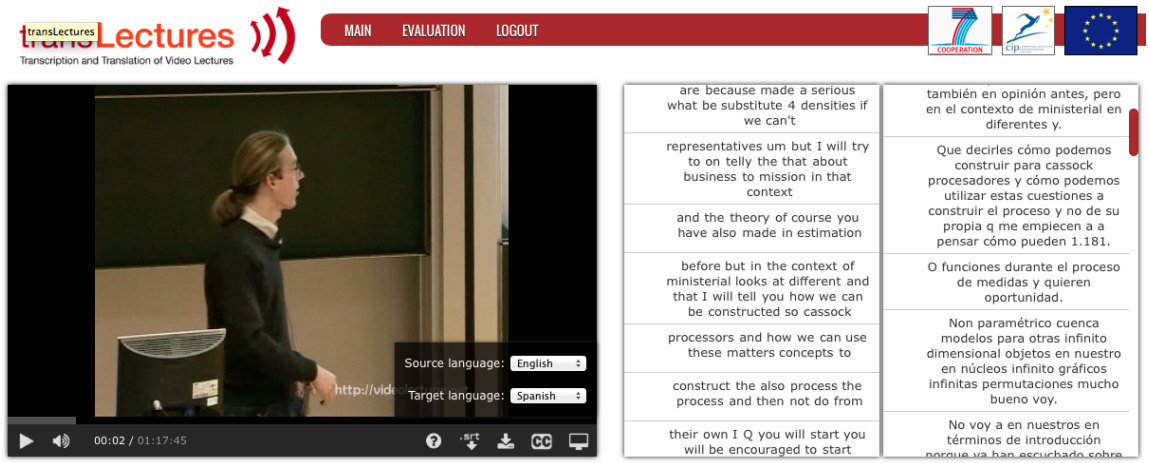

Figure 2: transLectures web player in translation mode for English into Spanish with the side-by-side layout.

2.2 Guidelines to be followed

It was agreed that a one-phase annotation would be used in the evaluation process, and transcription and translation guidelines were significantly simplified to account only for the time devoted to amend recognition and translation errors.

Transcriptions were automatically segmented before the supervision. The automatic segmentation process attempts to cut the transcription into natural fragments, so as to show at each point complete natural sentences or semantically complete fragments. It is also a goal to keep segments short enough for them to be easy to read and revise one at a time. To simplify the supervision process with a special focus on recognition errors, the automatic segmentation will not be modified by the experts.

The transcription guidelines used by the editors are defined in Appendix A.1, while the translation guidelines to be followed can be found in Appendix A.2. In addition, translation editors are to make annotations including a quality score at the sentence level in a scale from 1 to 5 (see Appendix A.3), as well as free text for any comments the editors might wish to make.

2.3 Selection of lectures

It was agreed that the selection of automatic transcriptions, to be manually supervised, would be made on the basis of transcription quality according to confidence measures at the video level. VideoLectures.NET lectures were mostly selected from the high confidence segment, while poliMedia lectures were selected according to the distribution of low, medium and high confidence lectures in poliMedia.

It was agreed that a total of 2 hours of program material per audio language (English, Spanish, Slovenian) would be used for manual supervision of transcriptions by language experts in this second round of evaluation. Once the transcription files were supervised, these same transcription files would be automatically translated into the target languages involved in transLectures. These automatic translations were the ones that were manually post edited. This means that the effort involved in the manual supervisions of automatic translations will be greater than that of the automatic transcriptions, as the translation language pairs (6 language pairs in total: English↔Slovenian, English↔Spanish, English→French, English→German) are more than the transcription languages (3 languages in total: English, Spanish, Slovenian).

The lectures selected for manual supervision in terms of automatic transcriptions and translations are shown in Appendix A.4.

3 Transcription Quality

In what follows, Sections 3.1, 3.2 and 3.3 analyse the transcription quality per each language of the lectures that were manually supervised in VideoLectures.NET and poliMedia, including summaries of the editors' comments and analyses of the automatic quality metrics. Then, Section 3.4 provides an overview of the transcription results achieved in this second quality control.

3.1 Quality control on English transcriptions (DDS)

There were 8 automatic transcriptions of lectures to be evaluated in total, 6 of which were of high confidence, 1 of medium and 1 of low confidence. Details of these lectures can be found in Table 11. The automatic transcriptions were generated by EML and were reviewed by two DDS editors.

The effort editors put in evaluating the files generally corroborated the confidence score of the transcription with one exception: Lecture #5140, of high confidence, was the file that took the longest to evaluate. The explanation for this discrepancy may lie both with the content of the lecture itself and with the editor selected to work on it. In brief, a different editor worked on the evaluation of this lecture to the editor that evaluated all the rest of the files. This difference in evaluation speed might have to do with a difference in the ability of the two editors to evaluate at a faster or lower speed in general, or with the fact that the editor of Lecture #5140 (Editor 2) had to get used to the functionality of the transLectures player itself, which he found to be rather slow as compared to software he is accustomed to using. When asked about the content of the lecture, the editor reported that he had to make changes in almost every single line, which was time consuming, even if the change was a single word only, but that the speaker of the lecture spoke clearly enough, albeit off-the-cuff at times, and without a strong accent.

In terms of the rest of the lectures, our editor (Editor 1) reported that the more general the topic, the better he found the ASR output to be. Terminology was generally misrecognised in all lectures, so the ones that were highly specialised required a higher supervision effort also. An

example of this is Lecture #5692, which abounded in chemistry terms and made the supervision fairly slow. The two easiest lectures to evaluate were #4774 and #4775, both on environment and relatively terminologically-free.

The medium confidence lecture (#12071) had quite a few transcription errors, as the speaker had a strong Spanish accent and fairly bad English in terms of grammar, which made it harder to work on the file, but on the other hand, the speaker was very clear in terms of what he was talking about, which made it more straightforward and quick to recognise and correct the mistakes.

The low confidence lecture (#15684) was the worst Editor 1 worked on. The speaker was talking fairly quickly, but his speech was conversational, so not too difficult to understand. However the echo in the audio quality must have made a difference to the ability of the recogniser to properly transcribe his speech, as the transcription quality in this lecture was poor in comparison to the rest this editor evaluated.

Both editors felt that the quality of the automated transcriptions was much better that those they evaluated in the first round, and that they were able to supervise them much faster than the ones last year. However neither of them could really pinpoint how much the availability of an automated transcription would have affected their productivity, as the transLectures player that they were asked to use for the evaluations is not a tool they had used before or would normally use and, hence, could not make a comparison to what their usual output would have been. When asked for a guesstimate, the editors reported time savings of at least 50% as compared to transcribing from scratch, which might even go up to 60%-70% at times. They generally felt that their work correcting the transcriptions involved much fewer keystrokes than having to transcribe everything from scratch themselves due to the high number of words that correctly transcribed automatically.

The automatic metrics WER[1]) and RTF[2] for these lectures are provided in Table 1. It is expected that lower WER figures correlate with less editor effort in terms of RTF.

Table 1: Summary of WER figures on supervised English lectures.

Id	Confidence	WER	RTF
15684	Low	28.7	7.5
12071	Medium	42.1	4.5
15073	High	27.5	6.1
5692	High	27.1	5.4
5393	High	23.3	4.9
5140	High	19.2	6.2
4774	High	15.2	4.7
4775	High	10.4	3.0
Total		23.9	5.2

If we now compare the editors' comments above to the WER of their files, we see that the editors subjective evaluation of the effort involved in their task generally corroborates the WER findings, with a couple of deviations: The low confidence lecture (#15684) has a lower WER than the medium confidence one (#12071), though the editor felt the effort he had to put on

[1]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. The lower the WER, the higher the transcription quality.

[2]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. The lower the RTF, the less effort invested in supervision.

the low confidence lecture was more than in the medium confidence one. Lecture #5140 that was evaluated by Editor 2 had a very low WER score, despite the higher effort this editor put in as compared with that of his colleague on other lectures. As previously stated, however, this could be the result of a difference in speed between the editors themselves and also due to the fact that Editor 1 had been accustomed to using the transLectures player by the time he evaluated the high confidence lectures, whereas his colleague hadn't.

In summary, the quality of English automated transcriptions in this round of evaluations was much better than the ones supervised last year; so much so that their availability could have produced significant time-savings in a commercial use-case scenario. For this to be corroborated with exact numbers though, experiments with transcription instructions that would include the addition of punctuation marks to the files (which were not to be used in this round of evaluations) would be needed. Also, the issue that slowed our editors down the most was having to research terms and proper nouns online to confirm their correct spellings, which might not have been necessary for someone specialising in each lecture's topic.

3.2 Quality control on Slovenian transcriptions (DDS)

There were 8 automatic transcriptions of lectures evaluated in total, 6 of which were of high confidence, 1 of medium and 1 of low confidence. Details of these lectures can be found in Table 12. The automatic transcriptions were generated by EML and a single editor from DDS evaluated all the files.

The effort our editor put in evaluating the files generally corroborated the confidence scores of the automatic transcriptions.

Indeed our editor found the low confidence lecture (#12363) to be the worst in terms of quality – by far. The editor reported that the bulk of the text had to be re-transcribed from scratch. The reasons for this had to do with the speaker himself: he was standing far from the microphone, used a colloquial Ljubljana accent which means most of the word endings were cut off, did not articulate very well, was speaking off-the-cuff quite often and the speech was not very coherent overall.

The medium confidence lecture (#3724) was problematic mostly because it was terminologically rich. The speech was highly specialised (on physics) and our editor had to research all of the terms mentioned, which she was not familiar with, to verify whether they were transcribed correctly or not. She managed to find a good chunk of the actual text online, which helped her verify that most of the terminology had indeed been transcribed correctly in the first place. Obviously someone specialising in physics would not have needed to spend as much effort in supervising this lecture as our editor did.

The two best lectures according to our editor were #14905 and #2078. In lecture #14905, the speaker was a Serbian woman, who spoke Slovene with a Serbian accent but very clearly. She articulates perfectly and talks as if from a book, paying particular attention to her grammar and sentence construction. In lecture #2078 the speaker actually read from a book. She was standing relatively far from the microphone and would occasionally cut words off, but on the whole she spoke very articulately. The differentiating factor in this speech was that it was not a free speech, which is probably why it was so clear and less problematic to transcribe automatically. Our editor reports that these two lectures were the ones where she was able to evaluate the fastest, as the bulk of the automatic transcriptions were near perfect.

Lecture #14030 was another lecture that was not a free speech, but where the speaker read from a script most of the time. This speaker also had a Ljubljana accent, and several words that were cut off as a result were also misrecognised, but on the whole the transcription was quite accurate.

Lecture #9797 was a TV show on chemistry involving three speakers, the presenter and two invited guests. When there was no overlap between the speakers the transcription was very accurate. Problems arose during speaker overlap, with the recogniser not being able to detect properly who was speaking and the automatic transcriptions not making sense as a unit.

Lecture #8563 was a speech about a book on climate change. The speaker in this case speaks as if her speech is a direct translation from English at times, making her Slovene sentence structure slightly awkward. The transcription was quite good however, most of the terms were also recognised correctly, but this was not the case with the English words present in the speech.

Lecture #12916 was the most problematic and time consuming one out of the high confidence lectures. The speaker speaks very fast and also has a Ljubljana accent. The speech is free, almost as if there are no sentence boundaries, so the editor needed to go back and forth in the speech repeatedly to make sense of what the speaker was saying. Foreign words in this lecture were also misrecognised. This was also the first lecture our editor evaluated, so the extra effort in evaluation might also have to do to an extent with the editor getting used to the transLectures player, that she hadn't used before.

Overall our editor reported that this set of transcriptions showed a significant improvement in quality as compared to the ones we supervised last year. A large number of transcriptions were near perfect and the editor reported a definite increase in productivity as a result of having the automatic transcriptions at hand. She guesstimates that this could have well been up to an 80% gain in productivity as compared to transcribing from scratch, but obviously specially designed experiments would be needed to corroborate such a productivity increase.

The automatic metrics WER and RTF for the Slovenian lectures are provided in Table 2 If we compare the editor's comments above to the WER of her files, we see that the editor's subjective evaluation of the effort involved in her task generally corroborates the WER findings.

Table 2: Summary of WER figures on supervised Slovenian lectures.

Id	Confidence	WER	RTF
12363	Low	82.2	10.4
3724	Medium	49.4	8.6
12916	High	25.0	7.5
9797	High	23.8	4.3
14030	High	21.5	4.3
8563	High	20.4	4.9
14905	High	19.8	3.5
2078	High	16.5	2.8
Total		31.6	5.4

In summary, the quality of the Slovene transcriptions supervised in this second round of evaluations was significantly higher compared to those supervised in the first round. The productivity gain saving as a result of the automatic transcriptions is guesstimated to be up to 80% as compared to transcribing from scratch (not taking punctuation into account, which we did not deal with in this evaluation round) and our editor could not repeat often enough how impressed she was at the improvement in quality in this evaluation round.

3.3 Quality control on Spanish transcriptions (UPVLC)

A total of 2 hours of video lectures in Spanish from the poliMedia repository were selected for the manual supervision and evaluation of their automatic transcriptions. This selection was

made at three confidence-level sets accounting for 20 minutes of high confidence transcriptions (4 lectures), 1 hour and 25 minutes of medium confidence transcriptions (11 lectures) and 15 minutes of low confidence transcriptions (5 lectures). More details of this selection can be found in Appendix A.4.

3.3.1 Experimental setup

The experimental setup, in terms of transcription guidelines and user interface (as described in section 2), was designed to reproduce the conditions under which transcriptions are supervised by regular users of the transLectures system. Under these conditions, user interaction and timing statistics are collected to analyse the behaviour of the experts and their performance compared to transcribing from scratch.

3.3.2 Experimental results

Table 3 details lecture Id, confidence level, WER and RTF. The RTF has been computed from the statistics collected by the transLectures player, taking into account the time spent by the expert listening to the lecture and typing to correct the automatic transcription. The supervision times recorded by the transLectures player match those manually recorded by the experts.

Table 3: Lecture Id, confidence level, WER and RTF data on the selected set of supervised Spanish transcriptions.

Id	Confidence	WER	RTF
7218	Low	41.9	5.0
6720	Low	41.7	6.8
2430	Low	35.9	5.0
594	Low	32.3	3.6
6489	Low	29.9	5.5
8600	Medium	29.2	5.3
7002	Medium	26.8	4.1
168	Medium	26.3	3.4
6297	Medium	25.9	5.7
7073	Medium	22.5	3.0
1769	Medium	20.5	3.5
2300	Medium	20.3	5.1
9156	Medium	20.1	2.5
5013	Medium	17.5	2.8
6015	Medium	17.5	2.6
7351	Medium	11.5	2.7
5684	High	14.9	3.0
8501	High	12.7	2.0
7481	High	11.0	1.9
1829	High	9.6	2.0
Total		23.2	3.8

Regarding the WER data, the figures reported for this selected set of poliMedia lectures are slightly higher than those for the test set defined in Task 6.1 (D6.1.2 [2]), but in any case they are representative of the complexity of the poliMedia task.

As to the confidence measures estimated by the ASR system for each lecture, the results show that they correlate to the final WER and RTF figures. Low confidence transcriptions were the least accurate ones (their average WER was 37.1) and took longer to supervise (their average RTF was 5), medium confidence transcriptions fell in the middle in both respects (average WER of 22.4, average RTF of 3.9), while high confidence transcriptions were the most accurate ones (average WER of 12.1) and took less time to supervise (with an average RTF as low as 2.2).

If we compare these results to the ones reported in deliverable D6.2.1[1] last year, in which the average effort required to supervise automatic transcriptions was similar to transcribing them from scratch (RTF of 10, approximately), it is obvious that there has been a clear improvement in the effort required.

This time, all transcriptions required an RTF of less than 6 for their supervision, except for ones of the least accurate, which required a slightly greater effort (RTF of 6.8). In fact, most of the transcriptions (15 out of the 20) required an RTF less than or equal to 5.

It is also worth noting how the most accurate transcriptions, those with a WER lower than 20, required an effort of less than 3 RTF for their supervision. This group of 7 transcriptions (8 if we include one with an RTF of 20.1), coming from the high and medium confidence levels, can be clearly seen in the lower left corner of figure 3 (section 3.4).

This year's results for transcription supervision, thus, show a very significant reduction of effort, as reflected in the average RTF of 3.8, less than 50% with respect to the figures reported last year (an improvement that can be seen both globally and within each confidence level).

Two important changes from last year's quality control must be taken into account when considering these improvements in RTF: first, that the tool used for supervision changed from Transcriber to the transLectures player (see section 3.3.4 below); and second, that this year's transcription guidelines did not ask for the addition of extra annotations to the transcriptions, which reduces the effort required but actually brings the supervision task closer to the conditions under which supervisions are expected to be carried out in real applications according to the transLectures approach.

3.3.3 Discussion on automatic transcription quality

Regarding the disparity in automatic transcription quality (with WERs ranging from 41.9 to 9.6), the main factors observed in the video lectures and affecting quality were similar to the ones reported last year in D6.2.1.

Audio quality was not a determining factor, as all poliMedia video lectures are recorded in a dedicated recording studio and so audio quality is similarly high for most lectures (even for the ones in the low confidence set).

The main factors observed were still the speaker's speaking abilities and whether the lecture's contents had been previously prepared in detail or not. Lectures in which the speaker enunciates each word clearly, uses short and well-structured sentences and keeps a correct rhythm of speech (e.g., Lecture Id 1829, on the use of a web application, and Lecture Id 7481, on statistics) were regularly better transcribed than lectures in which the speaker speaks in a rushed manner, uses improvised sentences or tends to mispronounce words and to correct themselves (e.g., Lecture Id 6489, on office software, and Lecture Id 2430, on teaching).

The technicality of the language was not such an important factor in this case. Some lectures were of a more technical nature than others, but the amount of technical terms was relatively small overall. Lecture Id 7481, on statistics, is an example of a lecture of a more technical nature in which technical terms were generally well transcribed; while Lecture Id

6297, on electric engineering, is an example of a lecture where technical terms were frequently mistranscribed (possibly because the speaker was prone to pronouncing words rushedly).

As to loan words and foreign words, there were also relatively few in this representative sample of lectures. As was to be expected, sometimes the speakers pronounce them with their corresponding foreign language phonetics and sometimes with phonetics adapted to Spanish language, which can be an added challenge for the speech recognition system. Lecture Id 594, on computer networks, is an example of a lecture where the speaker uses several foreign words, showing both ways of pronouncing at different points; in this lecture, these terms were frequently mistranscribed.

Lastly, a note with respect to the recognition of lectures by speakers with American Spanish accents. In this selection there were in total 6 such lectures, with final WERs as varied as 11 (Lecture Id 7481, on statistics), 20.3, 22.6, 26.8, 32.3 and 41.9 (Lecture Id 7218, on educational technologies). So the recognition of American Spanish accents was not worse in general than that of standard European Spanish accents; transcription quality did not depend on the speaker's accent, but rather, again, on their clarity when speaking (e.g., the speaker's enunciation and rhythm were better in Lecture Id 7481 than in Lecture Id 7218).

3.3.3.1 Automatic segmentation and transcription readability

The automatic segmentation was not corrected by the experts. This sped up the revision process as compared to last year's quality control (described in D6.2.1), where transcriptions were manually segmented by the experts. But it also meant that at some points the resulting segments were not optimal.

The automatic segmentation produced, in general, subtitles which were easy enough to read and revise. But in many points the segments were not semantically complete, and so a more natural segmentation could have been achieved. Some simple manual corrections would have improved the overall quality of the segmentation, probably taking less time than manually segmenting each transcription from scratch.

While improving automatic segmentation is not within the main objectives of the transLectures project, it is an interesting future field of work which could help improve the naturalness and readability of automatic subtitles. The criteria used by professional subtitle editors has been used as reference so far, and there is still useful work to do in this direction (e.g., automatic paraphrasing to reduce the length of the subtitles, ensuring that each segment is a complete sentence or a semantically complete fragment, ensuring that segments are not too short and not too long for comfortable reading, etc.).

3.3.4 Feedback on transLectures player usability

The capabilities of the transLectures player as a video lecture player and transcription editor were useful to speed up the supervision process.

With the transLectures player, during the supervision of a given video lecture's transcription, the expert was able to check simultaneously the lecture's video and audio, the accompanying slides and the transcribed text, all of it coordinated, segment by segment, moving back and forth when necessary, while introducing corrections within the same interface.

For the expert, this meant having immediate access at each point to all of the information typically offered by a video lecture repository. Specifically, the information provided by the slides was very useful to correct the transcribed text when words were not clear enough in the audio (usually because of unclear or rushed pronunciation by the speakers, as recording quality was generally high).

3.4 Comparison of transcription quality results

In this section we compare transcription results for English and Slovenian in VideoLectures.NET and Spanish in poliMedia. Figure 3 confrons WER scores to RTF for the aforementioned languages. As observed, most transcriptions with WER scores below 20 were supervised in less than RTF 3, that is, more than 3 times faster than transcribing from scratch. For transcriptions with WER higher than 20, RTF figures are above 5 but below 7 in most cases for all languages involved. Another conclusion in Figure 3 is that the Spanish editor seems to be faster at supervising than the English and Slovenian editors, for transcriptions in the same WER range. It should be noted that the adjustment to the points in Figure 3 is linear, but the adjustment curve seems to be quadratic because the x axis is displayed in logarithmic scale.

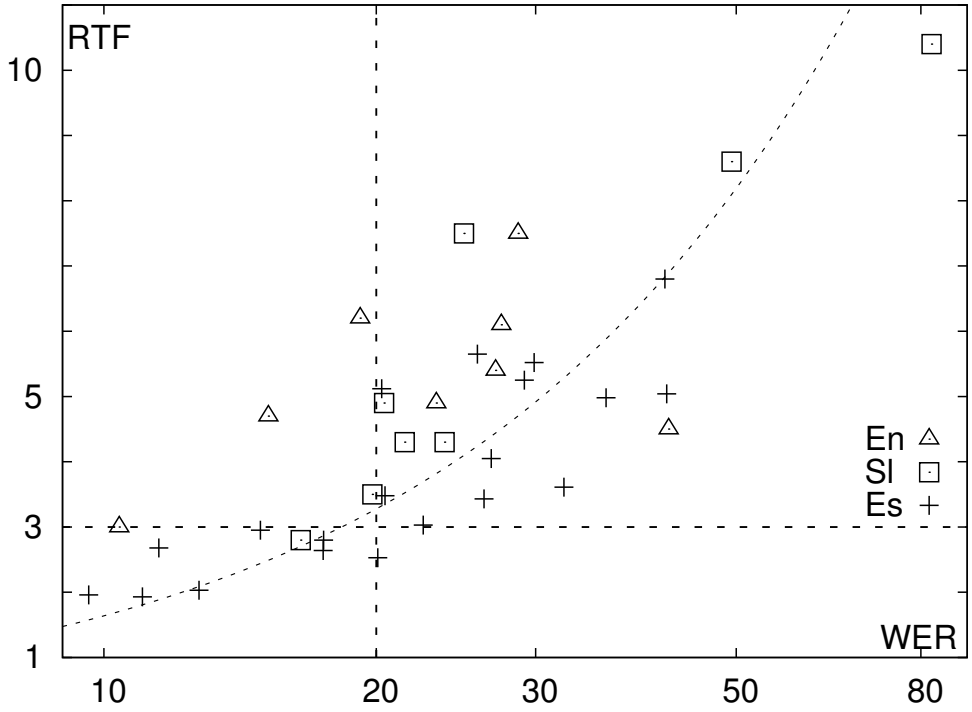

Figure 3: RTF versus WER (log scale) for English, Spanish and Slovenian transcription supervisions.

4 Translation Quality

What follows is an analysis of translation quality for the lectures whose automatic translations were manually post-edited.

In sections 4.1 to 4.6, a summary of the editors' comments is provided for each language pair, together with analyses of the human evaluation scores and automatic quality metrics. All these results for the six language pairs are then summarised and compared in section 4.7.

4.1 Quality control on Slovenian into English translations (DDS)

There were 8 automatic translations of lectures evaluated in total, provided by XRCE and evaluated by DDS. Details of these lectures can be found in Appendix A.4.

Although our editor reported that the quality of the Slovenian to English translations was better than that of the English into Slovenian ones, she was not pleased with their overall quality, as there was a lot of work involved in making the translations intelligible. The editor felt there was no improvement to the translation quality as compared to Year 1. When asked to rank the quality of the files on a scale from 1 to 5 (where 5=best), she ranked the overall quality as 2.5, i.e. below average, and reported that no lecture stood out in terms of quality. It is worth noting though that, by averaging out the quality score our editor allocated on each individual segment, the average subjective quality is average, i.e. 3, as shown in Table 4. Table4 gives, for each lecture, lecture Id, average manual score[3], Human BLEU[4] (HBLEU), Human TER[5] (HTER) and RTF[6]. The adjective *Human* refers to the fact that reference translations were obtained from amending automatic translations, so editors could be biased to obtain a reference translation that is closer to this generated automatically.

The editor believes it was the poor segmentation and lack of punctuation in the source files that caused serious translation issues, and this could have been the reason no notable improvement in quality could be seen in comparison to Year 1 evaluation where the source texts were segmented and punctuated properly. Mistranslations abounded (e.g. 'shoulder' instead of 'cutting') and it was often impossible to make use of the MT suggestions (e.g. 'there fundamental to ensure that' instead of 'there is no general assurance'). In order cases the wrong meaning of a word was used in the translation, so the translation was out of context or the text did not flow naturally (e.g. 'then' instead of 'actually'). The word order frequently had to be changed and this had to be done across segments too, as poor segmentation in the source files often meant that a sentence continued over more than one segments. Capitalisation did not seem to be taken into account and even simple things, such as the pronoun 'I', appeared in lowercase throughout the MT files and had to be corrected. Incorrect articles and prepositions were used (e.g. 'faculties in practice' instead of 'faculties to practice'), and proper names were in some cases literally translated, e.g. 'Meden' translated as 'honey' (Lecture #9797) and 'Stanovnik' as 'inmate' (Lecture #14030). The editor did report though that in these lectures researching and correcting terminology was not a time-consuming factor, as it was either translated correctly, or the Slovenian original that was left in could easily be modified to the correct English version, e.g. by a simple change in word suffixes.

[3]The 1-5 scale for the manual quality scores is outlined in Appendix A.3. The higher, the better.

[4]BLEU is an automatic quality metric designed to approximate human judgement. Its value is given as a number from 0 to 100, indicating how similar the candidate and reference texts are: the higher the number, the better the translation quality.

[5]Translation Error Rate (TER) is the ratio expressed as a percentage between the number of basic word editing operations to convert the automatic translation into the reference correct translation, and the number of words in the reference translation. The lower the TER, the higher the translation quality.

[6]In this case, RTF (real time factor) is the ratio between the time devoted to supervising the translation of a video and the duration of that video. The lower the RTF, the less effort invested in supervision.

When asked about the usefulness of the machine translations in a commercial setting, the editor reported that she did not think she would enjoy any productivity gain, and post-editing the files might even slow down her speed and cause a productivity loss. However, manual recordings of RTF indicate a productivity gain of approximately 50% as opposed to translating from scratch, as per Table 4.

It was noted that all Deluxe translators working on evaluating MT output seemed to think that their productivity would not be affected much by the availability of such output, as compared to translating from scratch. It is likely that this is because of the cognitive load translators have to deal with when engaged in the post-editing task that they are not accustomed to and need to familiarise themselves with. In the case of our Slovenian editor in particular, she had to work on a much larger volume of files that any other editor, as she dealt with the automatic transcriptions first, the automatic En>Sl translations next and finally the automatic Sl>En translations. By that point, she felt overwhelmed by the cognitive load of post-editing what she felt were below average machine translations and hence her perception of usefulness of the MT output might have been somewhat affected by this.

Table 4: Summary of manual evaluation metrics vs. automatic quality metrics in Slovenian into English translations.

Lecture ID	Score	HBLEU	HTER	RTF
8563	3.0	36.9	44.6	17.2
14905	3.2	35.5	44.8	14.4
3724	3.3	35.5	44.8	34.7
12916	2.9	33.5	49.9	24.5
14030	3.1	33.3	48.0	18.0
9797	3.1	32.8	46.6	18.7
12363	3.1	32.6	47.6	17.5
2078	3.0	30.4	51.1	15.0
Average	3.1	33.5	47.3	20.0

4.2 Quality control on English into Slovenian translations (DDS)

There were 8 automatic En>Sl translations of lectures to be evaluated in total, provided by RWTH and evaluated by Deluxe. Details of these lectures can be found in Appendix A.4.

Our editor was not happy with the quality of the files and she reported that they still seemed to be of the same low quality as in the Year 1 evaluation. Translation issues abounded and most of the text had to be retranslated from scratch, with extremely few segments being perfectly clear and intelligible. The editor suspects that this might have been in part due to the poor segmentation and lack of punctuation in the source files (transcriptions). When asked to rank the quality of the files on a scale from 1 to 5 (where 5=best), she ranked the overall quality as 2, i.e. poor, as shown in Table 5.

On the whole she found all files to be of equally poor quality, with Lecture #5692 standing out from the rest. This was the only lecture for which both the speaker and the topic were among the speakers and topics evaluated last year, so perhaps the difference in quality can be accounted for because of this.

One positive remark our editor made was that, although there was no punctuation in the source files, some commas did appear in the translations, particularly in Lecture #15684. When this happened, their positioning in the sentence was always correct.

Our editor guesstimates that using MT files of such quality would have slowed her down in a commercial use-case scenario and maybe even result in a productivity loss. The only lecture in which this wasn't the case was #5692, for which our editor guesstimates a 40% productivity gain. Automated and manual recordings of RTF though reveal an average productivity gain of roughly 30%, with this number going up to +60% in the case of Lecture #5692, as shown in Table 5.

It was noted that all Deluxe translators working on evaluating MT output seemed to think that their productivity would not be affected much by the availability of such output, as compared to translating from scratch. It is likely that this is because of the cognitive load translators have to deal with when engaged in the post-editing task that they are not accustomed to and need to familiarise themselves with.

Table 5: Summary of manual evaluation metrics vs. automatic quality metrics in English into Slovenian translations.

Lecture ID	Score	HBLEU	HTER	RTF
4775	2.4	24.7	53.6	20.4
12071	2.4	22.3	52.9	21.1
5692	2.4	18.9	55.8	15.1
15073	1.7	16.4	63.4	25.0
4774	1.6	14.2	64.8	23.5
5140	1.8	13.5	65.2	30.2
15684	2.2	13.1	65.4	28.8
5393	2.2	11.6	67.4	22.4
Average	2.1	16.4	61.5	23.3

4.3 Quality control on English into German translations (DDS)

There were 8 automatic En>De translations of lectures to be evaluated in total, provided by RWTH and evaluated by Deluxe. Details of these lectures can be found in Appendix A.4.

Our editor was not happy with the MT output in any of the lectures evaluated, as there were far too many errors and the editor proceeded with re-translating most of the text. The most recurrent, time-consuming and frustrating errors had to do with sentence structure and word order, which was simply wrong in all lectures. Other recurrent issues had to do mistraslations, either in the case of homonyms (e.g. the phrase 't sub k' in Lecture #12071, which was constantly mistranslated as t u-boat k' despite the fact that context could have indicated what the correct translation was) and particularly with short words, such as 'so', 'now', 'also', that can have different translations depending on context, but also use of non-German words (typically French, Slovenian and Russian words, e.g. climatis, divljacima, etc) instead of the proper German translation.

The editor thinks that a lot of the translation issues stem from the missing punctuation and less than perfect segmentation of the source texts (transcriptions). She pointed out that, had she been provided the English transcriptions alone, without the videos, it would have been hard for her to understand what the lectures were about. It was also hard for the editor to not use punctuation or capitalisation in the translation task and this made it problematic to offer a more accurate opinion on the quality of the output, which she rates as poor (2 on a 1-5 quality scale where 5=best).

The two lectures that stood out for her were Lecture #5393, which was particularly challenging as it was full of terminology and was very hard for her to understand in the first place, not

being an expert. The other was Lecture #12071, which was very simple and only had very few terms that needed researching, but which were repeated throughout the text and consistently mistranslated.

The editor does not think the current MT output would have saved her any time as compared to translating from scratch, and that at most it would have helped her with word choice in some cases. Our assumption is that translation from scratch is a x40 RTF, although this is not likely to be so in this editor's case, as she is a particularly fast translator. In any case, manually recorded RTF is roughly x13, as per Table 6, which still indicates added efficiency using the MT output than working from scratch.

Furthermore, it was noted that all Deluxe translators working on evaluating MT output seemed to think that their productivity would not be affected much by the availability of such output, as compared to translating from scratch. It is likely that this is because of the cognitive load translators have to deal with when engaged in the post-editing task that they are not accustomed to and need to familiarise themselves with.

Table 6: Summary of manual evaluation metrics vs. automatic quality metrics in English into German translations.

Lecture ID	Score	HBLEU	HTER	RTF
4775	2.6	21.4	65.4	8.7
12071	2.5	20.4	75.6	14.8
15073	2.0	17.4	70.6	13.2
15684	2.1	15.5	64.3	20.3
5692	1.9	15.2	72.7	11.9
5393	2.0	15.2	74.1	12.5
5140	2.1	13.7	72.7	9.7
4774	1.7	13.6	76.3	11.5
Average	2.0	16.3	72.4	12.8

4.4 Quality control on English into French translations (DDS)

There were 8 automatic En>Fr translations of lectures to be evaluated in total, provided by XRCE and evaluated by Deluxe. Details of these lectures can be found in Appendix A.4.

In general, our editor found the quality of the files to be average to poor (2.5 on a 1-5 scale where 5=best) and reported the same types of errors as those found in the Year 1 Evaluation: concordance, conjugation, syntax, word order, OOV, mistranslations, etc. Very few segments were found to be perfectly clear and intelligible, whereas most were fraught with grammar and syntax issues. Word order was particularly problematic to fix and, where terminology was involved, this typically required a lot of research. Interestingly, when estimating the average quality score the translator allocated during the evaluation on a segment by segment basis, as per Table 7, the average subjective quality is 3.5 instead, i.e. above average.

Some of the most recurrent errors included the below:

- Contracted forms, such as 'll or 're, were almost systematically left as such in the MT output and needed to be removed.
- Words with hyphens were often left in English.
- Proper nouns were not capitalised.

- Many verbs appeared in the infinitive form when they should have been conjugated.

Our editor did not seem to think that any of the lectures stood out particularly in terms of quality. When asked to rank the lectures from best to worst quality-wise, the following ranking order was provided:

1. Lecture #5692: Lecture 36: Review
2. Lecture #4775: New solutions for collective transport: fuelling bus rapid transit (BRT) in Europe
3. Lecture #12071: Triangular Numbers (Part III)
4. Lecture #4774: Achieving Sustainable Urban Mobility: The Challenges
5. Lectures #15073 (Customising the multilingual customer experience ? deliver targeted online information based on geography, user preferences, channel and visitor demographics) #5140 (Seedcamp)
6. Lecture #5393: Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression

In particular, Lecture #5393 was deemed the worst mainly because it contained a lot of specialised terminology that needed to be thoroughly researched, which was very time-consuming. In general, researching terminology was found to be one of the issues that slowed down editors the most, as they had to spend time to research it, irrespective of whether it had been properly translated or not.

On the whole, our editor reported that, with the amount of work involved in correcting the MT output, she could not see how this would help increase her productivity in a commercial use-case scenario. However, manual recordings of RTF indicate that the post-editing process seems to have sped up our translator's work by roughly 50% on average.

It was noted that all Deluxe translators working on evaluating MT output seemed to think that their productivity would not be affected much by the availability of such output, as compared to translating from scratch. It is likely that this is because of the cognitive load translators have to deal with when engaged in the post-editing task that they are not accustomed to and need to familiarise themselves with.

Table 7: Summary of manual evaluation metrics vs. automatic quality metrics in English into French translations.

Lecture ID	Score	HBLEU	HTER	RTF
12071	4.0	45.4	34.2	16.7
4775	3.9	42.7	42.0	15.3
5393	3.4	37.1	48.3	19.7
5692	3.7	34.7	50.8	16.5
4774	3.4	33.2	55.7	16.8
15073	3.3	26.2	60.7	18.8
15684	3.6	23.6	58.6	25.0
5140	3.2	22.5	64.1	18.9
Average	3.5	33.4	52.6	18.5

4.5 Quality control on English into Spanish translations (DDS)

There were 8 automatic En>Es translations of lectures to be evaluated in total, provided by UPVLC and evaluated by Deluxe. Details of these lectures can be found in Appendix A.4.

Our editor found the quality of the translations to be fairly good (4 on a 1-5 scale where 5=best) and thought the MT output was quite accurate and acceptable. When estimating the average quality score the translator allocated to the files during the evaluation on a segment by segment basis, as per Table 8, the average subjective quality is slightly lower, i.e. 3.6 or above average.

The recurrent errors found in the texts were of all types, the most frequent ones being similar to what was noticed in other language pairs:

- Word order
- Agreement / conjugation
- Formal/informal treatment
- Missing articles / wrong prepositions
- OOV
- Words that are not present in the original that needed to be deleted
- Numbers that did not match
- Passives used when not frequent in Spanish
- Words with multiple meanings translated wrongly, e.g.
  - Using “por” instead of “para” for “for” (Lecture #12071)
  - “Since” is not always “desde”, also “ya que” (Lecture #12071) or sometimes “como” (Lecture #4775)
  - “Now” translated as “ahora” when it is used as a word to change subject (Lecture #15684)
  - “Then” is always translated as “luego”, it can be “despus”, “entonces” (Lecture #12071)
  - “We” translated as “nos” when it is “nosotros” (Lecture #5692)
  - “A lot of” is translated as “un montn de” but that is too colloquial (Lecture #5692)
  - “Answer” can be verb or a noun, always translated as a verb (Lecture #4775)
  - “That” always translated as “que” when it is a pronoun (should be “eso”) (Lecture #4774)
  - “Where” not only means “donde” but also “aunque” (Lecture #5140)
  - Confusion with verb to be in Spanish (ser/estar) (Lecture #15073)

Our editor did not find any of the lectures to differ in translation quality, it was rather the topic difficulty that differed. As the lectures were all specialised, she needed to ensure she understood the original first before post-editing the MT output. She pointed out that in some cases it was the source text that did not make much sense, as speakers were not finishing their sentences (e.g. Lecture #15073), therefore the MT output was inevitably problematic. There were also problems with the translation of fillers, e.g. ‘so’, ‘now’, etc. that are not used as frequently in Spanish and there is not a big variety of them either. So she had to use the word

'vale' repeatedly to translate these fillers, when in many cases a Spanish speaker would simply say 'eh' a lot instead.

She also warned that there is a risk that translators, when presented with MT output, might be influenced by it and that they should be provided with appropriate training so that they do not blindly accept the options provided to them. When asked to rank the lectures in terms of translation quality, the below ranking order was provided:

1. Lecture #5140: Seedcamp
2. Lecture #15684: Introduction to the Workshop on Online Trading of Exploration and Exploitation 2
3. Lecture #12071: Triangular Numbers (Part III)
4. Lecture #4774: Achieving Sustainable Urban Mobility: The Challenges
5. Lecture #4775: New solutions for collective transport: fuelling bus rapid transit (BRT) in Europe
6. Lecture #5692: Lecture 36: Review
7. Lecture #15073: Customising the multilingual customer experience ? deliver targeted online information based on geography, user preferences, channel and visitor demographics
8. Lecture #5393: Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression

On the whole the editor was fairly pleased with the MT output and reported that this would definitely provide a productivity boost to her in a commercial use case scenario. She guesstimates this to be 40%. The manual recording of RTF indicates even higher productivity savings, as high as 70%, as shown in Table 8.

It was noted that all Deluxe translators working on evaluating MT output seemed to think that their productivity would not be affected much by the availability of such output, as compared to translating from scratch. It is likely that this is because of the cognitive load translators have to deal with when engaged in the post-editing task that they are not accustomed to and need to familiarise themselves with.

Table 8: Summary of manual evaluation metrics vs. automatic quality metrics in English into Spanish translations.

Lecture ID	Score	HBLEU	HTER	RTF
4775	3.9	59.9	26.0	11.8
12071	3.6	58.1	26.0	8.6
4774	3.6	56.9	31.7	12.0
5393	3.5	52.1	32.7	14.2
15073	3.3	49.7	35.2	16.4
5692	3.6	49.0	33.9	10.3
5140	3.4	43.8	39.5	16.1
15684	4.1	42.6	40.3	11.3
Average	3.6	51.4	33.3	12.6

4.6 Quality control on Spanish into English translations (UPVLC)

A total of 2 hours of automatically translated poliMedia video lectures were manually edited (supervised) and evaluated. These Es>En translations were generated at the UPVLC from the corrected Spanish transcriptions resulting from the manual supervision process described in Section 3.3. Table 13 with details of the 20 lectures evaluated can be found in Appendix A.4. Human evaluation scores, automatic quality metrics, and editors' comments are summarised below.

4.6.1 Experimental setup

The experimental setup, including translation guidelines and the user interface (as described in Section 2), was designed to reproduce the conditions under which translations are supervised by regular users of the transLectures system. Under these conditions, user interaction and timing statistics are collected to analyse the behaviour of the experts and their performance compared to translating from scratch (from the original transcription).

4.6.2 Experimental results

Table 9 gives, for each lecture, lecture Id, average manual score, HBLEU, HTER and RTF.

The RTF indicates the times manually recorded by the experts performing the supervision. For transcription (Section 3.3), the transLectures player was able to accurately record the total supervision time for each video lecture as time spent listening plus time spent typing corrections. In the case of translation supervision, however, it is not so straightforward: in order to carry out this task, the human editors frequently consulted information sources outside of the player, which introduces an additional time factor that is not so easily measured.

As shown in the table, the average RTF for translation supervision was 14.8. This means that the process of editing one hour of transLectures translated subtitles would take 14.8 hours to complete. This result is a significant improvement on previous figures of 34-40 RTF for the process of translating a video lecture from scratch (from its transcription) and means that, for this set of lectures, experts were able to produce correct translations more than 2.25 times faster than if translating from scratch.

Regarding the manual quality score, the average quality score of 3.6 (out of 5) compares very favourably to the 2.9 awarded by the same experts in the first quality control. This improvement in the quality of the automatic translations as perceived by the experts is greater than might have been expected considering the more moderate (although significative) improvement in average HBLEU from 42.4 to 46.6.

In Table 10, we can appreciate the distribution of the scores awarded to each segment in the translations. Comparing to the results reported last year, the distribution has changed from a dominance of the poorer scores 1 and 2 (which have decreased from summing 42% of the segments to their current sum of 20.5%) to a dominance of the higher scores 4 and 5 (which have grown from their previous sum of 34% to currently summing 54% of the segments).

4.6.3 Discussion of quality and editor commentary

The quality of the automatic translations in this evaluation set varied from lecture to lecture, with HBLEUs ranging from 29.1 to 68.4, and HTERs from 30.1 to 16.0.

As reported last year, a key factor impacting the quality of the automatic translations is the competence of the speaker when presenting their subject. This has much to do with the

Table 9: Summary of manual and automatic quality metrics on the selected set of Spanish into English translations.

Lecture ID	Score	HBLEU	HTER	RTF
8600	3.8	68.4	20.5	14.7
5684	3.7	56.4	26.4	9.2
7073	3.8	55.9	32.4	9.9
7351	4.2	53.6	30.5	7.7
9156	3.2	51.7	25.9	11.1
1769	3.9	51.6	31.6	20.0
7481	3.5	50.5	35.2	15.4
1829	3.6	49.4	31.7	7.4
6489	3.9	48.8	38.1	5.0
7218	3.6	47.9	40.2	15.3
594	3.2	47.6	35.6	11.6
2300	3.7	46.2	34.5	18.4
6720	3.5	46.2	38.2	15.0
6297	3.6	45.3	38.5	18.9
6015	3.5	43.5	42.2	15.2
5013	3.4	43.0	39.3	9.5
8501	3.5	40.2	42.0	11.2
168	3.5	39.1	40.0	18.5
2430	3.3	33.6	36.3	16.0
7002	2.8	28.9	46.6	18.8
Total	3.6	46.6	36.4	14.8

Table 10: Human score summary: Average percentage of segments awarded each manual quality score (from 1 to 5), total number of segments, and average manual score on the selected set set of Spanish into English translations.

1	2	3	4	5	Segments	Avg. Score
5.5%	15.0%	25.5%	26.5%	27.5%	1399	3.6

extent to which the lecturer prepared their talk before turning up to the recording studio: the more 'scripted' the speech, the better the automatic translation. The impressions of the experts in this respect were in line with that reported in D6.2.1: lectures in which the speaker used complete sentences, with complete grammar structures and more standard language resulted in higher-scoring translations and, consequently, faster RTFs. Lectures which had more incomplete sentences, repetitions or self-corrections, and more colloquial turns of phrase tended to result in lower-scoring translations.

Some frequently occurring mistakes which were identified this time round are recurrent from those reported in D6.2.1. Again as suggested in the first quality control, many of them might be easy to avoid or solve via an automatic post-processing stage:

- The automatic translations contain a mix of American English (AE) and British English (BE) spellings, often within the same lecture (e.g. characterizes/characterises, color/colour).
- Contractions are used incoherently (e.g. "we're" and "we are" appear in the same lecture).

Meanwhile, one expert noted a distinct reduction in frequency of other common mistakes reported in D6.2.1, such as noun and adjective pairing appearing in reverse order.

Another issue relates to segmentation, which often saw sentences divided across two or more segments. We discuss this in more detail below, in 4.6.3.1. We also discuss the possible impact of the new transcription and translation guidelines used in this quality control, in 4.6.3.2 and 4.6.3.3.

In any case, this round of quality control has quantified the impact of implementing the transLectures interactive player for translating video lecture transcriptions. The numbers, especially compared to last year's non-interactive method which saw no real improvement in RTF, speak for themselves.

Alternative implementations and setups could be considered to address some of the issues discussed below. For instance, the transLectures player might be modified to allow manual re-segmentation.

4.6.3.1 On the automatic segmentation

The segmentation of the automatic translation was inherited from the automatic segmentation of the transcriptions; that is, the transcription was translated segment by segment and, since this segmentation was not manually edited at the transcription stage, it was often less than optimal. Segments that bore no relation to what would constitute a natural or semantically-correct fragments or sentences were therefore carried over into the translation.

In the first instance, it seems clear that this interfered with the automatic translation process itself; automatic translations techniques were applied to segments in isolation from the rest of their natural semantic context. Segments that contained the end of one sentence and the start of the next were treated as a normal sentence. Similarly, given that word order can often differ between Spanish and English, having a sentence span over two or more segments leads to, at best, unnatural-sounding translations.

Secondly, given that this difference in word order, the correction process was far more cumbersome, with editors having to move back and forth between segments to rearrange the parts of the target text. In fact, it would be fair to say that, even when comparative word order was not an issue, the fact that single sentences were broken up over several lines impaired the readability of the translation. This, in turn, made it harder to evaluate and correct the translation.

A related concern might be that any inter-segment corrections of the kind discussed above would have repercussions on the data used to retrain the translation/language models.

All this said, revising the automatic segmentation in the original transcription prior to the automatic translation process, and/or allowing re-segmentation at the translation stage, would adequately address these matters.

4.6.3.2 On the transcription guidelines

One minor/infrequent translation error had its origin in the guidelines used to edit the original transcription before applying automatic translation: transcriptions were to match the phonetics of the speaker even when pronounced something incorrectly. This resulted in misspelled words being left in the transcription that were more problematic for the automatic translation system to recognise (e.g. (in Spanish) "red blackbone", instead of "red backbone", which was translated as "blackbone network" instead of "backbone network").

Another transcription guideline was that transcriptions be unpunctuated, which led to the automatic translation system introducing punctuation incorrectly into the translations[7].

4.6.3.3 On the translation guidelines

The focus of this round of translation quality control was on measuring how fast editors could correct the mistakes present in the automatic translation using the transLectures player in order to produce understandable subtitles. To this end, editors were asked simply to correct the automatic translation with a view to producing a meaningful translation of the transcription, repetitions and all. They were asked to ignore elements of punctuation, including capitalisation, and spelling. AE and BE were both accepted in the automatic transcription; as were correct capitalisations and incorrect non-capitalisations; and translated and non-translated names of universities, for instance.

The reason for this was that the transLectures player and intelligent interaction are intended for minor corrections of near-perfect translations, and not re-translations. And our editor's comments do confirm that the transLectures system is "built for speed"; it encouraged a faster pace and more minor corrective effort, and they experienced less temptation to make non-strictly-necessary or stylistic changes. However, combined with the difficulties of the automatic segmentation mentioned above, the two above points mean that some subjective concern could be raised regarding the end quality of the translations obtained following this protocol, which might be considered fit-for-purpose rather than of professional quality.

4.6.4 Feedback on transLectures player usability

The capabilities of the transLectures player as a video lecture player and transcription editor were useful to speed up the translation supervision process.

With the transLectures player, during the supervision of a given video lecture's translation, the expert was able to check simultaneously not only the lecture's transcription and translation side by side, but also the lecture's video and audio and the accompanying slides, all of it coordinated, segment by segment, moving back and forth when necessary, while introducing corrections within the same interface.

The experts' impression was that the supervision process using the transLectures player felt generally faster than last year (D6.2.1), when translation supervision was instead carried

[7]We tried to avoid this affecting automatic metrics, by asking that the editors neither add, move nor remove the automatic punctuation. However, this was a more troublesome request than it seem, and it has been suggested that this might have led to higher RTFs because the editors had to consciously ignore these sentence markers.

out having the transcription and the translation side by side, and manually playing the corresponding audio in specific cases of doubt. This impression is corroborated by the measured RTFs.

According to the experts, having the video and audio at hand and synchronised to the texts helped enormously, especially since they are spoken texts, and so disfluencies, repetitions, unfinished sentences, etc. are easier to understand and translate when heard.

4.7 Comparison of translation quality results

In this section we compare translation results across the different language pairs in poliMedia and VideoLectures.NET. Figure 4 shows HBLEU scores versus RTF for the six translation language pairs. As can be observed, the supervision of the translations from English into Slovenian presenting the lowest quality required far more time than the translations from English into the other languages. However, the translations into German, even though they have a similar quality compared to the ones into Slovenian, were supervised in less time by an exceptionally fast expert. Next, reading the plot from left to right, the translations into French and into Slovenian come next with RTF values below 20 in most cases. Finally, the translations from English into Spanish (VideoLectures.NET) and from Spanish into English (poliMedia) obtain the highest HBLEU scores, and present RTF figures below 20 in all cases, even below 10 in some cases.

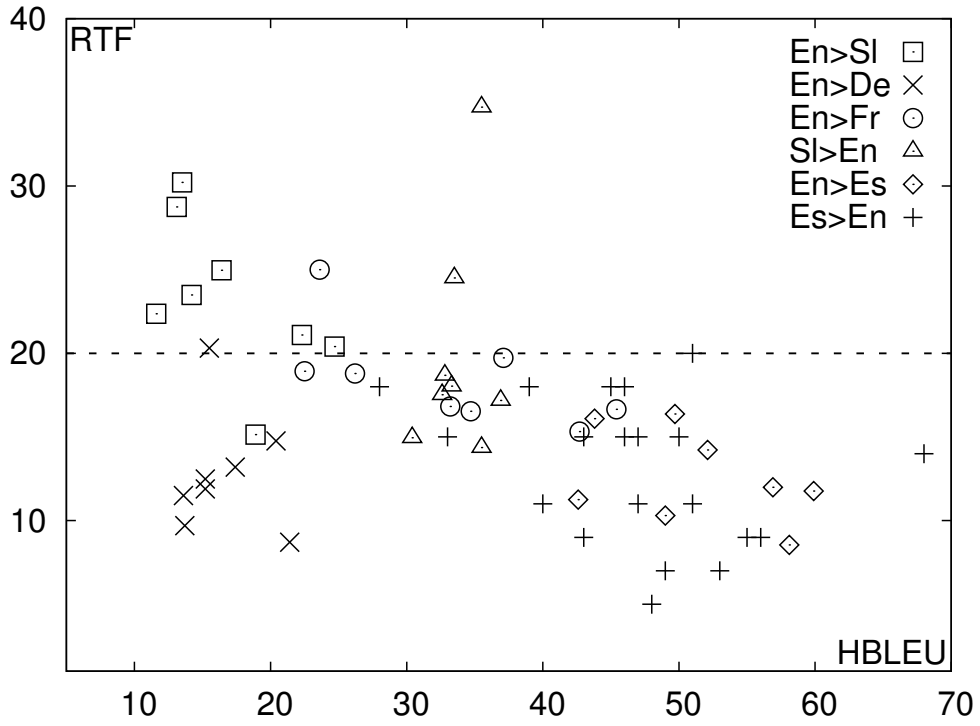

Figure 4: RTF versus BLEU for all translation pairs.

Figure 5 shows HBLEU scores versus the manual quality scores (HSCR). The translations from English into Slovenian and into German have a HSCR below 3, which correlates with their HBLEU scores below 30. Then, from left to right, the translations into French and the translations from Slovenian into English obtain higher HSCR scores as their HBLEU scores increase. Finally, as in Figure 4, the translations from English into Spanish (VideoLectures.NET) and from Spanish into English (poliMedia) show only minor HSCR improvements compared to the ones we just mentioned, even with significantly higher HBLEU scores.

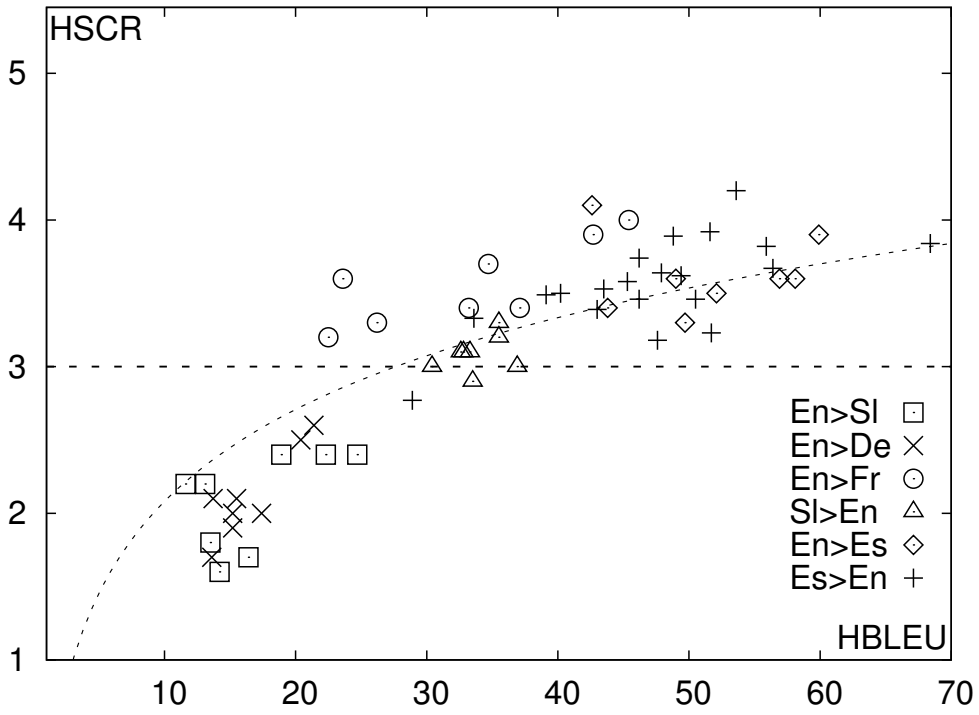

Figure 5: Quality score versus BLEU for all translation pairs.

5 Conclusions

By way of conclusion we can say that improvements were observed in both case studies. However, the extent of this improvement and the general impression of the experts seemed to depend on both the task, language and case study in question.

Summing up, for VideoLectures.NET, transcription quality was above average and all editors in all languages (En, Es, Sl) reported a significant increase in the quality of the transcription files they received relative to Y1. The recommendation in this case to achieve better RTFs would be the use of subject experts for the manual supervision process. The lack of annotations in this round of evaluation and the fact that automatic segmentation and segment timings were not to be corrected both helped decrease overall supervision time. Editors confirmed a definite increase in productivity in a commercial use-case scenario and guesstimated an average 50% productivity gain, which is also corroborated by the automatically and manually recorded RTF, and which would be significantly higher if only high confidence transcriptions are taken into account.

For translation in VideoLectures.NET, there was greater variation between languages, reflecting the same disparities as observed in Y1: En>Es scored highly, En>Fr, Sl>En scored as average, and En-De and En-Sl scored poorly. Accordingly, only the Spanish experts (so far) felt that using the MT output would lead to productivity gains. This could be because of the cognitive load translators have to deal with when post-editing, which is not a task they are accustomed to, and the lack of having a real measure as to how long translation from scratch would take them, as the instructions they were following in these evaluations are not the typical instructions translators would follow for work they complete in a commercial setting. Translators in this evaluation round agreed that an extra factor affecting MT quality output is the automatic (and often wrong) segmentation and lack of punctuation or capitalisation.

For poliMedia, the tests carried out on transcription quality revealed a direct correlation between WER and RTF which, although not unexpected in itself, validated the usefulness of

the confidence measures. On the whole, an average reduction of effort (RTF) of over 50% was recorded (making a conservative calculation). The same significant improvements were observed in translation: an average RTF reduction of over 50% was recorded relative to transcribing from scratch.

On the whole, we are optimistic that these productivity gains can be sustained and improved upon as project technologies, including the transLectures player, are refined over the coming months. It is worth noting in this respect that, as per the experts' commentary in earlier sections, although the transLectures system, in its current state, might be less than useful in some professional use-case scenarios (specifically, some of the professional translation scenarios), it is proving its worth in the context for which the tools are being designed: for use by video authors or video users, in the specific context of video repository annotation. Video authors, for instance, could certainly be considered as subject experts.

For its part, the transLectures player has been well-received. However, the experts expressed some dissatisfaction regarding the lack of some possible functionalities, a key issue being that it does not allow re-segmentation. Discussions are under way regarding the possibility of introducing a second expert mode where this (and other functions) would be possible, while keeping in mind that the main intended users are non-professionals.

References

[1] UPVLC, XEROX, JSI, RWTH, EML, and DDS. D6.2.1: First report on quality control. Technical report, transLectures, 2012.

[2] UPVLC, XEROX, JSI, RWTH, EML, and DDS. D6.1.2: Second report on scientific evaluations. Technical report, transLectures, 2013.

A Appendix

A.1 Transcription guidelines

1. Always transcribe what was said.
   - Transcribe phonetically correctly.
   - Always transcribe what is spoken, do not change or correct what the speaker says.
   - Do not correct grammar.
   - Do not capitalise at the beginning of a sentence.
   - Do not include punctuation marks.
   - Do not transcribe incomprehensible passages, leave them blank.
   - Leave out sounds such as "em, hm, eh", as casual users will not add these kinds of hesitations to a transcript in later real-life situations.
   - If a punctuation mark is spoken, it should be transcribed as a word.
     Examples:
     - "gonna": transcribe "gonna"
     - When the speaker corrects himself: transcribe what was said.
2. Letters and spelled out words: each letter will be individually typed in upper-case, separated from the other letters by a space. Example: my name is Smith S M I T H
   - For acronyms (e.g., USA), the individual letters will not be separated by spaces.
3. Numbers, digits, ordinals, variable names, equations:
   - Numbers, digits and ordinals will be transcribed as words. Examples: four hundred and eighteen, forty-fourth, two thousand and two
   - Decimals will be transcribed as words. Example: two point six
   - Variables will be transcribed as words. Examples: epsilon, lambda, ...
   - Equations must be written out in words as they are spoken.
     Example:
     - Speaker says: ... P equals N over V times R T ...
     - Correct transcription: ... P equals N over V times R T ...
     - Incorrect transcription: P=(n/V)(RT)
   - Zero (0), when pronounced "oh", must be transcribed as "oh"
   - Further examples: three squared plus two, seven point three minus three point eight
4. Incomprehensible passages are to be left blank.
5. The following non-speech events will not be transcribed:
   - Hesitations (voiced sounds): note that only non-word hesitations must not transcribed.
   - Speaker noise (unvoiced sounds): lip smack, etc.
   - Non-stationary noise (door slam, window, ...).
   - Music.
   - Chatter and noise.
6. Speech disfluencies:

- Incorrect pronunciation: imagine the speaker utters something incorrectly. Incorrect utterances are to be transcribed for all speakers!

  Example:

  - Speaker says: ... there was a 'prublem' with ...
  - Transcription: ... there was a prublem with ...
  - Explanation: The speaker pronounced 'prublem' but wanted to say 'problem'.
- Foreign words: for words, titles, etc. in foreign languages that are pronounced differently from English pronunciation rules (or Slovenian respectively), they are to be transcribed in the foreign language.

  Example:

  - Speaker says: ... there was a coup d'êtat in ...
  - Transcription: ... there was a coup d'êtat in ...
- Repetition: the speaker, unintentionally, repeats a word or expression. It will be transcribed as it is pronounced.
- Word cut-offs: imagine the speaker utters only half a word. This half-word is to be transcribed as spoken.

7. Overlapping speech: Transcribe only what the main speaker says.
8. Slovene lectures: Do not use correct grammatical spelling for colloquial terms; transcribe as spoken.

A.2 Translation guidelines

1. Follow transcription case, that is, lowercase.
2. Speech disfluencies:
   - Incorrect pronunciation: translate if understandable. Example:
     Transcription: ... there was a prublem with ...
     Translation: ... hubo un problema con ...
   - Foreign words: Leave untranslated. Example:
     Transcription: ...there was a coup d'êtat in ...
     Translation: ...hubo un coup d'êtat en
   - Repetition: Translate words repeated unintentionally.
3. Numbers, digits, etc: Follow the style in the transcriptions (i.e., full words if it's in full words in the source text, numbers if it's in numbers in the source text) for consistency.

A.3 Translation Quality Assessment Guidelines

For every sentence, please provide a single translation quality score, using a scale from 1 to 5, following the guidelines below. Deciding which score to attribute to a translation relates to both post-editing effort, as well as the gravity of the translation errors encountered.

For every point in the translation quality evaluation scale, a list of examples of translations from English to Spanish that were assigned this score in a similar evaluation is provided [**?**]. Each example provides the original English source sentence, the translation system's output that was evaluated and the result of human post-editing.

- Quality Score = 1

  The MT output is incomprehensible, with little or no information transferred accurately. It cannot be edited, needs to be translated from scratch.

  Example 1:

  English source:
  US artist's son 'in museum theft'
  Spanish translation system output:
  Nosotros, el hijo de artista "en museo robo"
  Spanish human post-edited:
  El hijo de un artista estadounidense involucrado en el robo en un museo

  Example 2:

  English source:
  Taxpayers meet mortgage on Tory's £75,000 orangery
  Translation system output in Spanish:
  Los contribuyentes reunirse con hipotecas sobre el tory 75 000 libras orangery
  Spanish human post-edited:
  Los contribuyentes pagan la hipoteca de 75.000 libras de un invernadero de naranjos de un tory

- Quality Score = 2

  About 50-70% of the MT output needs to be edited. It requires a significant editing effort in order to reach publishable level.

  Example 1:

  English source:
  Young bears, captured as they swam after the bodies of their newly killed mothers, were caught and sent to zoos and circuses.
  Translation system output in Spanish:
  Es joven, capturados como lo swam después de los cuerpos de sus recién mató a las madres, fueron capturados y enviado a los parques zoológicos y circo.
  Spanish human post-edited:
  Los cachorros fueron capturados cuando nadaban tras los cuerpos de su madres a las que acababan de matar, y los enviaron a parques zoológicos y circos.

Example 2:

English source:
The mature infectious viral corpuscles would not occur if the HIV virus was not split by the HIV protease.
Translation system output in Spanish:
La madura viral infecciosa corpuscles no se produciría si el virus VIH no estaba dividido por el VIH protease.
Spanish human post-edited:
Los corpúsculos virales infeccionsos maduros no se producirían si el virus VIH no fuse dividido por la proteasa del VIH.

- Quality Score = 3

  About 25-50% of the MT output needs to be edited. It contains different errors and mistranslations that need to be corrected.

  Example 1:

  English source:
  On a day when little happened in the U.N.-sponsored climate talks, thousands of activists walked across the city holding banners in English saying "There is No Planet B" and one in Spanish declaring, "The Earth is Saying, Enough."
  Translation system output in Spanish:
  En un día en que poco ocurrió en el clima U.N.-sponsored conversaciones, miles de activistas recorrieron en toda la ciudad sostener pancartas en inglés diciendo "No es No Planeta B" y uno en español declarar, "La Tierra está diciendo, Basta."
  Spanish human post-edited:
  En un día en el que poco ocurrió en las charlas sobre el clima patrocinadas por la ONU, miles de activistas recorrieron toda la ciudad llevando pancartas en inglés que decían "No hay un planeta B" y una en español que decía, "La Tierra dice, Basta."

  Example 2:

  English source:
  The Cabinet Office official who left top-secret documents on a train in June is to be charged under the Official Secrets Act, the BBC has learned.
  Translation system output in Spanish:
  El Gabinete funcionario que dejó arriba documentos secretos en un tren en junio va a ser acusado bajo el Official Secret Act, la BBC ha aprendido.
  Spanish human post-edited:
  El funcionario de la oficina del gabinete que dejó documentos ultra secretos en un tren en junio será acusado bajo el Acto de Secretos Oficiales, se ha enterado la BBC.

- Quality Score = 4

  About 10-25% of the MT output needs to be edited. It is generally clear and intelligible.

  Example 1:

  English source:
  According to the representative, it is necessary to inform the widest range of people - Roma and non-Roma people as well - about these interrelationships.
  Translation system output in Spanish:
  Según el representante, es necesario informar al mayor número de personas romaníes y no romaníes personas así - sobre estas interrelaciones.
  Spanish human post-edited:
  Según el representante, es necesario informar al mayor número de personas - de Roma y también fuera de Roma - acerca de estas interrelaciones.

Example 2:

English source:
AIG must quickly find the funds to repay the loan of 85 billion that it has with the Federal Reserve or else risk seeing the U.S. authorities take 80% of its shares, diluting the proportion due to current shareholders.
Translation system output in Spanish:
AIG debe encontrar rápidamente los fondos para pagar el préstamo de 85 millones de euros que tiene con la Reserva Federal o correr el riesgo de que las autoridades estadounidenses tienen el 80% de sus acciones, diluir la proporción debido a los actuales accionistas.
Spanish human post-edited:
AIG debe encontrar rápidamente los fondos para pagar el préstamo de 85 millones que tiene con la Reserva Federal, o de lo contrario arriesgarse a ver que las autoridades de los EE.UU tomen el 80% de sus acciones, diluyendo la proporción debido a los accionistas actuales.

- Quality Score = 5

  The MT output is perfectly clear and intelligible. It is not necessarily a perfect translation, but requires little to no editing.

  Example 1:

  English source:
  According to AFP, the gunmen are part of the New People's Army (NPA), which is the armed faction of the Communist Party of the Philippines (CPP).
  Translation system output in Spanish:
  Según AFP, los pistoleros forman parte del Nuevo Ejército Popular (NPA), que es la facción armada del Partido Comunista de Filipinas (CPP).
  Spanish human post-edited:
  Según la AFP, los pistoleros forman parte del Nuevo Ejército Popular (NPA), que es la facción armada del Partido Comunista de Filipinas (CPP).

  Example 2:

  English source:
  "Sooner or later it will end in a fatal accident," he told Týden.cz nearly a year ago.
  Translation system output in Spanish:
  "Tarde o temprano acabará en un accidente fatal", dijo Týden.cz hace casi un año.
  Spanish human post-edited:
  "Tarde o temprano acabará en un accidente fatal", dijo él a Týden.cz hace casi un año.

A.4 Lectures selected for quality control

Table 11: English lectures from VideoLectures.NET selected for quality control.

Id	Category	Confidence	Duration
15684	Events	Low	00:03:12
12071	Mathematics	Medium	00:09:01
15073	Marketing	High	00:18:38
5692	Chemistry	High	00:31:27
5393	Text Mining	High	00:24:20
5140	Venture Capital	High	00:12:41
4774	Environment	High	00:16:21
4775	Environment	High	00:12:24
Total			02:08:04

Table 12: Slovenian lectures from VideoLectures.NET selected for quality control.

Id	Category	Confidence	Duration
12363	Sustainable Development	Low	00:14:37
3724	Physics	Medium	00:12:22
12916	Innovation	High	00:07:26
9797	Materials	High	00:21:44
14030	Small and medium enterprises	High	00:12:52
8563	Politics	High	00:18:38
14905	Public health	High	00:08:55
2078	Teaching	High	00:24:05
Total			02:00:39

Table 13: Spanish lectures from poliMedia selected for quality control.

Id	Category	Confidence	Duration
7218	Educational technologies	Low	00:07:11
6720	Microsoft Office	Low	00:02:00
2430	Teaching	Low	00:01:34
594	Computer netwoks	Low	00:04:19
6489	Microsoft Office	Low	00:03:59
8600	Fine arts	Medium	00:02:23
7002	History	Medium	00:09:35
168	Information technologies	Medium	00:06:30
6297	Renewable energy	Medium	00:18:14
7073	Politics	Medium	00:06:33
1769	Managment	Medium	00:13:44
2300	Educational technologies	Medium	00:04:04
9156	Mathematics	Medium	00:06:46
5013	Educational technologies	Medium	00:05:47
6015	Sport management	Medium	00:05:35
7351	Design	Medium	00:03:14
5684	Autocad	High	00:03:48
8501	Marketing	High	00:05:48
7481	Statistics	High	00:06:29
1829	Teaching	High	00:04:03
Total			02:01:36

B Acronyms

AE	American English
ASR	Automatic Speech Recognition
BE	British English
BLEU	Bilingual Evaluation Understudy
DDS	Deluxe Digital Studios Limited
EML	European Media Laboratory GmbH
HBLEU	Human Bilingual Evaluation Understudy
HTER	Human Translation Error Rate
JSI	Josef Stefan Institute
K4A	Knowledge for All Foundation
MT	Machine Translation
RWTH	RWTH Aachen University
RTF	Real Time Factor
TER	Translation Error Rate
UPVLC	Universitat Politècnica de València
WER	Word Error Rate
XRCE	XEROX Research Center Europe