D2.3: Final report on complete transcriptions and translations

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

Distribution: Public

Project funded by the European Community under the Seventh Framework Programme for Research and Technological Development.
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 Media Europe (DDS)

For copies of reports, updates on project activities and other transLectures related infor-
mation, contact:

The transLectures Project Co-ordinator
Alfons Juan, Universitat Politècnica de València
Camí de Vera s/n, 46022 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

© 2014, The Individual Authors

No part of this document may be reproduced or transmitted in any form, or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission from the copyright owner.
1 Introduction

This deliverable reports on the transcriptions and translations created in WP2 'Maintenance of current transcriptions and translations’. The goal of WP2 was to maintain during transLectures a complete set of current transcriptions and translations. In the following a short summary of the work done to implement this is given.

In Task 2.1 a substantial amount of English and Slovenian lectures from VideoLectures.NET and Spanish lectures from poliMedia was identified and extracted for further processing. In M12 these lectures were automatically transcribed and translated into the following languages:

- **English lectures** were translated to French, German, Slovene and Spanish,
- **Slovene lectures** were translated into English,
- **Spanish lectures** were translated into English.

The work in WP3 'Massive Adaptation' was aiming at demonstrating that acceptable transcriptions and translations can be obtained by massive adaptation of acoustic, language and translation models from lecture-specific material (for details see [11, 14, 17]). Every 6 month the best models available from WP3 were tuned for applying it for the transcription and translation of thousands of hours of lectures. In month 18, 24, 30 and 36 the lectures were re-transcribed and then re-translated.

The transcriptions and translations were further improved by the methods developed in WP4 'Intelligent interaction with users'. Truly interactive models in which the system learns immediately from the interaction and different interaction modes corresponding to multiple user roles were developed. For details see [12, 15, 18]. At the case study sites different techniques were evaluated during internal and external evaluations organised in WP6 'Evaluation'. These evaluations also contributed to improve the transcription and translation quality of the repositories by providing manually reviewed subtitles. Details on the evaluations can be found in [16, 20].

For storage and retrieval of the transcriptions and translations repositories were implemented at the case study sites VideoLectures.NET and poliMedia. The implementation is based on the 'git' distributed revision control system. The automatically generated transcriptions and translations were committed to the respective repositories. For evaluation and quality analysis the transcriptions and translations were retrieved from the repositories and respective corrections were committed. Finally, the case-study websites retrieve transcriptions and translations from the video lecture repositories to be displayed to their visitants. A detailed description of the repository can be found in [13].

To store the transcriptions and translations the 'Timed Text Markup Language’ (TTML) format was adopted. The TTML is a content type that represents timed text media. To reflect all needs of transLectures an extension of the format was defined. The extension is described in [13].

The transcriptions and translations from M12, M18 and M24 are described in [13] whereas in this deliverable the focus is on transcriptions and translations generated in M30 and M36.

\[1\text{See http://www.w3.org/TR/ttafl-dfxp}\]
2 Automatic Transcription of Videolectures

In M12 initial automatic transcripts for English and Slovenian lectures from VideoLectures.NET and Spanish lectures from poliMedia were produced by EML.

In M18 UPV took over the transcription of the Spanish lectures and produced improved transcriptions in M18, M24, M30 and M36. In M24 UPV started to transcribe additionally Catalan lectures and in M30 English lectures available in the poliMedia repository.

EML continued to produce improved English and Slovenian transcriptions in M18, M24, M30 and M36 for the videolectures in the VideoLectures.NET repository.

2.1 Automatic Transcription of Videolectures from VideoLectures.NET

From VideoLectures.NET a set of more than 9,000 English lectures - corresponding to about 5,900 hours of speech - was identified based on the popularity (meaning the lectures that were viewed at least 50 times).

As for English the most popular Slovenian lectures were identified resulting in a set of about 800 lectures - corresponding to about 490 hours of speech.

From these two sets, lectures were separated for test and development sets, for quality analysis and several rounds of evaluation. Of course the lectures that were transcribed or corrected manually were excluded from further re-transcription.

Finally, in M36, a new set of more recent and untranscribed English lectures was transcribed for the first time by UPV. This new set comprised 935 lectures (628 hours) published in the VideoLectures.NET site between 8th July 2013 and 29th April 2014, and their transcriptions along with their translations into Spanish (also generated by UPV) were committed into the git repository in order to make them available on the VideoLectures.NET site. Complementarily, EML will process untranscribed video lectures from VideoLectures.NET after the project ends with the goal of having the transcriptions for all lectures recorded since transLectures started.

2.1.1 Automatic Transcription of English Videolectures

For the re-transcriptions acoustic and language models developed by EML or RWTH were used. The transcriptions in M30 were done using an AM and LM from EML: a TANDEM acoustic model was trained and a MAP adaptation with domain specific data was applied. The language model was improved by adapting with data coming from 3 new sources:

- Web material downloaded for the titles of the lectures.
- The manually corrected transcripts of the 2nd phase of evaluation.
- The automatic transcriptions from M24.

The final transcriptions in M36 were done with an AM and LM developed by RWTH that was ported and tuned for the EML Transcription Platform. Details on the systems used for the transcription can be found in [11, 14, 17]. Table 1 provides word error rates (WER) results on the test set as well as the average real-time factor (RTF) for the English transcriptions beginning in M12 until the final transcription in M36.
Table 1: WER and average RTF for the English transcriptions over transLectures.

<table>
<thead>
<tr>
<th>System</th>
<th>WER</th>
<th>RTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>M12 (EML)</td>
<td>44.0</td>
<td>8.5</td>
</tr>
<tr>
<td>M24 (RWTH)</td>
<td>31.5</td>
<td>3.1</td>
</tr>
<tr>
<td>M30 (EML)</td>
<td>28.2</td>
<td>8.0</td>
</tr>
<tr>
<td>M36 (RWTH)</td>
<td>23.4</td>
<td>5.2</td>
</tr>
</tbody>
</table>

Table 2: WER and average RTF for the Slovenian transcriptions over transLectures.

<table>
<thead>
<tr>
<th>System</th>
<th>WER</th>
<th>RTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>M12</td>
<td>47.5</td>
<td>1.5</td>
</tr>
<tr>
<td>M24</td>
<td>45.4</td>
<td>2.6</td>
</tr>
<tr>
<td>M30</td>
<td>40.5</td>
<td>7.0</td>
</tr>
<tr>
<td>M36</td>
<td>36.1</td>
<td>5.4</td>
</tr>
</tbody>
</table>

2.1.2 Automatic Transcription of Slovenian Videolectures

For the transcriptions acoustic and language models developed by RWTH were used. The models were ported and tuned for the EML Transcription Platform. Details on the systems used for the transcription can be found in [11, 14, 17].

Table 2 provides WER results as well as RTF for the Slovenian transcriptions beginning in M12 until the final transcription in M36.

2.2 Automatic Transcription of Videolectures from poliMedia

Since M18, UPV took full control of the live poliMedia repository. From the viewpoint of Task 2.3, this means that all lectures (in Spanish, Catalan and English) at poliMedia were automatically transcribed by the UPV. It is important to note that the poliMedia repository is continuously growing and new lectures have been added since the project started.

For this reason, at M18 and M24 the UPV transLectures team synchronised its local copy of poliMedia before each complete re-transcription was performed, but after the complete deployment and integration of the transLectures Platform (TLP) [19] with the poliMedia repository at M30, new recordings are being automatically transcribed and translated once they are published in the poliMedia catalogue. So far, since transLectures started approximately 5,000 new lectures (650 hours) have been incorporated into the poliMedia repository. Approximately 3,300 lectures (500 hours) were automatically transcribed and the rest of them were manually processed. Obviously, only automatic transcriptions are re-transcribed, while manual and supervised transcriptions are left out of this re-transcription process. Nowadays, automatic transcriptions account for 93% of the total amount of transcriptions in poliMedia.

Table 3 shows a brief summary of the number of poliMedia lectures that have been automatically transcribed for each language in M36 using the corresponding WP2 ASR system.

2.2.1 Automatic Transcription of Spanish Videolectures

In M36, from the 10,155 Spanish lectures available in the poliMedia repository, 8,610 lectures (1,147 hours) have been automatically transcribed given that the remaining 1,545 have been
Table 3: Number of lectures automatically transcribed by the UPV at M36 in each of the three languages considered for transcription in poliMedia, along with the number of hours and speakers that those lectures account for.

<table>
<thead>
<tr>
<th></th>
<th>Lectures</th>
<th>Hours</th>
<th>Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>8,610</td>
<td>1,147</td>
<td>1,494</td>
</tr>
<tr>
<td>English</td>
<td>564</td>
<td>80</td>
<td>256</td>
</tr>
<tr>
<td>Catalan</td>
<td>132</td>
<td>16</td>
<td>60</td>
</tr>
</tbody>
</table>

either manually transcribed (train, development and test sets) or manually supervised in the internal and external evaluations thanks to the full integration of transLectures into the poliMedia repository. Eligible lectures were automatically re-transcribed with the best WP3 ASR Spanish UPV system [17] but adapted to the WP2 requirements.

Firstly, all manually or supervised transcriptions from the poliMedia repository, except for the train, development, and test sets defined in Task 6.1, have been added to the acoustic training data of the ASR system, accounting for 50 additional hours of in-domain speech data with respect to the corresponding WP3 ASR system.

Secondly, due to the computational cost of the automatic transcription process, the ASR system has been tuned to reduce the time needed to re-generate the complete set of transcriptions for poliMedia. First, pruning parameters were optimised on the development set (defined in Task 6.1) in order to increase the transcription speed in terms of RTF without significantly compromising the transcription quality of the system. Last, language model (LM) adaptation had to be applied at the speaker level, instead of at the lecture level as performed in WP3. As a result, 1,494 adapted LMs (one per speaker) were trained, instead of 8,610 LMs (one per video).

Finally, the impact of the tuning process described above was evaluated in terms of WER on the development and test sets defined in Task 6.1. At M36, the WP2 system attained 13.6% and 14.4% on the development and test sets, respectively, while the WP3 system achieved 12.4% and 12.2% on the respective sets. We can notice a small gap between WP2 and WP3 systems quantifiable in 1.2% and 2.2% points on the development and test sets respectively, which can be explained by the pruning parameters adopted by the WP2 system, as well as the different scope of language model adaptation and the usage of an automatic speech segmentation system to detect speech segments, which in some cases might not be so accurate.

A summary of the progress of the Spanish transcription task can be seen on the left-hand side of Figure [1] which shows the progress over time of the WP2 Spanish ASR system in terms of WER computed over the test set defined in Task 6.1 for poliMedia, in comparison to the Spanish ASR systems reported at WP3 by UPV (WP3U) and RWTH (WP3R). Regarding the WP2 curve, it must be noted that in M12 the system used in the initial transcription of the whole poliMedia repository was provided by RWTH and hosted at the EML Transcription Platform (R@E).

2.2.2 Automatic Transcription of Catalan Videolectures

Catalan lectures available at the poliMedia repository were also transcribed at M30 and M36. Currently, the total number of Catalan lectures at poliMedia is 340, accounting for 42 hours, from which 187 were manually transcribed to define train, development and test sets, and 21 have been manually supervised in the Y3 internal and external evaluations thanks to the full integration of transLectures into poliMedia. Therefore, the remaining 132 lectures (16 hours) were automatically transcribed with the best WP3 ASR Catalan UPV system, which scores 17.4 WER points on the Catalan test set [17].
2.2.3 Automatic Transcription of English Videolectures

At M30 the UPV started to transcribe English lectures in poliMedia. At this moment, the poliMedia repository offers 626 English lectures which sum up to 91 hours, from which 57 lectures were manually supervised in the Y3 internal evaluations. Then, the remaining 569 lectures (83 hours) were automatically transcribed at M36 with the best WP3 ASR English UPV system with 21.4 WER points on the VideoLectures.NET English test set [17].

3 Automatic Translation of Videolectures

Transcriptions available at the VideoLectures.NET and poliMedia repositories were then translated. RWTH was performing the translations from English to German and Slovene, XEROX from English to French and Slovene to English and UPV from English to Spanish and vice versa. Furthermore, UPV started to translate Spanish lectures into Catalan and vice versa and Catalan lectures into English and vice versa. So, poliMedia video lectures are available with Spanish, English and Catalan subtitles.

3.1 Automatic Translations of Videolectures from VideoLectures.NET

3.1.1 English→German

For translation of the English VideoLectures.NET database into German, we applied the state-of-the-art decoder from RWTH’s open source machine translation toolkit Jane [21]. The phrase-based system is trained on all available data and is augmented with sophisticated sentence selection and weighting techniques to adapt it towards the lecture domain [7, 6, 5]. We use a 4-gram language model trained with the SRILM toolkit [10]. The standard models including phrase translation probabilities and lexical smoothing in both directions, word and phrase
penalty, an n-gram target language model, distance-based reordering model and three binary count features are applied. The log-linear parameter weights are optimized with MERT \cite{8}. We leveraged large amounts of in-domain and out-of-domain data to train a system with the best possible performance. Compared to the last period, the system was improved with a hierarchical lexicalized reordering model \cite{3} and 7-gram word class language models \cite{22} and maximum expected BLEU training. A more detailed description of the translation system can be found in \cite{17}. The neural network models mentioned there could not be used for WP2 due to their computational complexity. Translating the entire VideoLectures.NET repository required 8G memory and 198.7 days CPU time, corresponding to a speed of 2.65 words per second.

3.1.2 English→Slovenian

The English→Slovenian translation system is very similar to the one used for the English→German task. The same open-source decoder is used with the same basic features. The baseline system makes use of all available bilingual and monolingual data sources to achieve state-of-the-art performance and is further improved by adapting it towards the lecture domain with data selection, weighting techniques and discriminative training. Due to improved reference translations, we were able to improve translation results by 4% BLEU absolute. The translation system is described in detail in \cite{17}. To translate the entire VideoLectures.NET repository, the system required a maximum of 8.8G memory and 216.4 days CPU time, which corresponds to a speed of 2.43 words per second.

3.1.3 English→French

The MT system used in the English→French translation task is based on the open-source toolkit Moses \cite{4}. The decoding process uses the standard phrase-based approach. This system is domain adapted to the task and uses approaches developed during Month 24 to Month 36. Namely, they are the relevance features, the language models array and the meta-parameter optimization techniques (see deliverable \cite{14}). Target language models are trained with the SRILM toolkit \cite{10} and then binarized.

The system used is the best one created by Xerox and described in detail in the final report on massive adaptation \cite{17}.

3.1.4 English→Spanish

In order to translate VideoLectures.NET English lectures into Spanish, the UPVLC has used the best English→Spanish WP3 system described in the final report on massive adaptation \cite{17}, which is based on the open-source phrase-based state-of-the-art Moses system \cite{4}, and consists on an interpolation of the translation and language model cross-entropy techniques described in \cite{17}. This system replaced the previous M24 system which was based on the bilingual sentence selection (BSS) technique, a very expensive approach in computational terms. Therefore, M30 and M36 English→Spanish translations could be produced up to 3 times faster in comparison with the previous BSS-based systems. This translation system achieved 35.6 BLEU points in the VideoLectures.NET English→Spanish test set.

3.1.5 Slovenian→English

For Month 36, in the Slovenian→English translation task, Xerox used the best system presented in the deliverable \cite{17}. The same approaches were used as in Section 3.1.3 for the English→French task plus the reverse translation approach. This approach consists in the use of the translations from English to Slovene from Month 30, provided by our partner RWTH, to increase the amount of training data. The data and methods are detailed in the final report on massive adaptation \cite{17}.
3.2 Automatic Translations of Videolectures from poliMedia

The UPV generated in M30 and M36 automatic transcriptions of Spanish, Catalan and English poliMedia lectures. For each transcription, automatic translations have been generated in the other two poliMedia languages to ensure that every lecture provides subtitles in the three poliMedia languages.

3.2.1 Spanish→English

In order to translate Spanish poliMedia lectures into English, the UPV has used the best WP3 Spanish→English MT system described in the final report on massive adaptation [17], which is based on the open-source phrase-based state-of-the-art Moses system [4], and consists on an interpolation of translation and language cross-entropy techniques described in [17]. As in the English to Spanish system described in Section 3.1.4, this replaced the previous M24 system which was based on the BSS technique. Therefore, as in the English to Spanish system, M30 and M36 Spanish→English translations were generated up to 3 times faster in comparison with the previous BSS-based system. Finally, this translation system achieved 28.2 BLEU points in the poliMedia Spanish→English test set.

A summary of the progress of the Spanish→English translation task can be viewed at the right-most plot of Figure 1, which shows the progress over time of the WP2 Spanish→English MT systems in terms of BLEU computed over the test set defined in Task 6.1 for poliMedia, in comparison to the corresponding system reported in the WP3.

3.2.2 English→Spanish

English→Spanish of poliMedia lectures were generated using the English→Spanish MT System for VideoLectures.NET which is described in Section 3.1.4.

3.2.3 {Spanish, English}→Catalan and Catalan→{Spanish, English}

Translations from and to Catalan of poliMedia lectures were generated using the open-source rule-based translation system Apertium [2]. Apertium yields high quality Spanish to Catalan (and vice versa) translations, since Spanish and Catalan are very similar languages in lexical, syntactical and grammatical terms. Acceptable translations from Catalan into English and vice versa are provided by Apertium, despite the problem with out-of-vocabulary words. However, little human effort is required to generate these translations, that is, the built-in Apertium decoder needs to be executed with the text to be translated as input.

4 Speech synthesis of video lectures

Following the reviewers’ recommendation, the UPVLC-transLectures team has started to tackle the problem of automatically synthesizing a voice track from the subtitles of a lecture. This procedure will allow the users to enjoy the talks in their own language, letting them focus on the slides and not on the subtitles.

With this purpose, from M24 onwards the UPVLC team has carried out the development of two synthesis systems, one for Spanish and another for English. These systems are based
on statistical parametric speech synthesis \cite{24}, and make use of the latest state-of-the-art techniques, like HNM vocoding and neural network-based acoustic modeling \cite{11, 23}. Acoustic model training was performed with a modified version of the TLK recognition toolkit, and the audio samples were extracted from a subset of videos from the poliMedia and VideoLectures.NET repositories. In order to enforce temporal alignment with the video, the synthesized voice is synchronized at the start of every subtitle segment, and the duration is lengthened or shortened accordingly. Preliminary results of this work have been published in \cite{9}.

With the intention of performing a subjective evaluation of the systems, the supervised Spanish to English translations produced during the life of this project have been synthesized. This corpora is comprised of 140 lectures from the poliMedia repository, and corresponds to the development and test subcorpora of the poliMedia English system, the lectures from the Y1 and Y2 quality control performed by DDS, as well as some extra talks pertaining to the most prolific poliMedia lecturers. Additionally, lectures from the first two Spanish courses available on the European MOOC aggregator EMMA have been also synthesized in English and will be available soon. In total, over 31 hours of speech have been processed and synthesized.

Finally, the \texttt{transLectures} player has been modified with the addition of an extra button to allow external audio streams to be played, synchronized with the video. This extension enables the deployment of new or improved synthetic voices without needing to modify the lecture file.

5 Conclusion

In this deliverable, we have presented the work done to maintain current transcriptions and translations for the case study sites VideoLectures.NET and poliMedia during the project.

In the first year the most popular lectures from VideoLectures.NET and all lectures from poliMedia were indentified for transcription and translation throughout the project, resulting in about 5,900 hours of English, 490 hours of Slovenian and 700 hours of Spanish lectures.

In M12 initial transcriptions and translations for these lectures were produced with the models developed in WP3 ‘Massive Adaptation’. To store the transcriptions and translations the ‘Timed Text Markup Language’ (TTML) format was extended based on the needs of \texttt{transLectures}. At each case study site a repository was implemented to store all transcriptions and translations.

All lectures were retranscribed every 6 month using the improved models developed in WP3. To apply transcription technology for mass-decoding of thousands of hours of lectures resource (CPU time and memory) consumption has to be restricted and the developed models were tuned to balance accuracy and real-time factor.

Due to the dynamic nature of these repositories, UPV decided to update its local copy of poliMedia at M18 and M24 in order to transcribe and translate all lectures not included at the beginning of the project. Starting from M30, the poliMedia repository is uploading all newly recorded videos to the \texttt{transLectures} server once they are added to their catalogue thanks to the complete deployment and integration of the \texttt{transLectures} Platform (TLP) \cite{19} into the poliMedia repository. In M24 UPV additionally started to transcribe and translate Catalan lectures from poliMedia, and in M30, English lectures as well.

In addition, internal and external evaluations at WP6 contributed to provide manually supervised subtitles from those automatically generated in WP2, improving the overall quality of transcriptions and translations in VideoLectures.NET and poliMedia.

Finally, as observed in Figure \ref{fig:transcription_quality}, the transcription quality was substantially improved throughout the project. WER figures for automatic transcriptions were decreased from M12 to M36 by 47\%, 24\%, 36\% and 51\% relative in English, Slovenian, Spanish and Catalan, respectively. Translation quality was in most cases that provided by the best WP3 systems due to the reduced computational cost of the translation processs compared to the transcription process.
Figure 2: Overall WP2 Progress throughout the transLectures project in ASR for English, Spanish, Slovenian and Catalan transcriptions in terms of WER, all of them computed over their respective test sets defined in T6.1.

References


