D6.2.1: First report on quality control

DDS, EML, XEROX, UPVLC and RWTH

Distribution: Public

trans Lectures
Transcription and Translation of Video Lectures
ICT Project 287755 Deliverable D6.2.1

October 31, 2012

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 (XRCE)
Josef Stefan Institute (JSI)
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

© 2012, 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.
Executive Summary

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

Contents

1 Introduction 4

2 Selection of Lectures and Assignment of Tasks 4

2.1 Tool for transcription and translation evaluations 4
2.2 Guidelines to be followed 4
2.3 Selection of lectures 5

3 Transcription Quality 5

3.1 Quality control on English transcriptions (DDS) 5
3.2 Quality control on Slovenian transcriptions (DDS) 6
3.3 Quality control on Spanish transcriptions (UPVLC) 8

4 Translation Quality 9

4.1 Quality control on Slovenian into English translations (DDS) 9
4.2 Quality control on English into Slovenian translations (DDS) 10
4.3 Quality control on English into German translations (DDS) 11
4.4 Quality control on English into French translations (DDS) 13
4.5 Quality control on English into Spanish translations (DDS) 13
4.6 Quality control on Spanish into English translations (UPVLC) 15

5 Conclusions 16

A Appendix 19

A.1 Transcription guidelines (DDS) 19
A.2 Translation guidelines (DDS) 22
A.3 Translation Quality Assessment Guidelines (XRCE) 23
A.4 Lectures selected for quality control (EML and UPVLC) 26

B Acronyms 28
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 first report on the quality of said transcriptions and translations.

Next section explains how lectures from VideoLectures.NET and poliMedia were selected for manual supervision. 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 set of selected lectures are included 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 to also be added to the current transcriptions and translations in WP2.

Discussions on the selection of the transcribed and translated lectures for manual evaluation begun among relevant project partners in August 2012. After thorough discussion, the below was agreed.

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 Transcriber, for the following reasons:

- Consistency with the work done in Task 2.2.
- Familiarity with Transcriber tool by DDS translators, so as to avoid re-training on a new tool.
- Familiarity with Transcriber tool by the rest of the project partners, all of who have the ability to convert to the .trs format using a script provided by UPVLC.

2.2 Guidelines to be followed

It was agreed that a one-phase annotation would be used in the evaluation process.

For transcriptions, the original transcription guidelines are to be used by the evaluators (see Appendix A.1). Editors are to amend incorrect timings of sentences and add notes in the form of Transcriber annotations for all sentences the timings of which have been altered as well as for any free text comments the editors wish to make. No questionnaire is to be used for transcription evaluations.

For translations, the original translation guidelines are to be followed (see Appendix A.2), with the exception of the use of hesitations and repetitions, which have already been eliminated in the automatic translations provided to the editors. Editors are to make notes in the form of Transcriber annotations per sentence. Annotations should include a quality score at sentence level in a scale 1-5 (see Appendix A.3), as well as free text for any comments the editors wish to make.
Table 1: Number of hours of transcription and translation selected in the first report on quality control for the languages involved: English (En), Spanish (Es), French (Fr), German (De) and Slovenian (Sl).

<table>
<thead>
<tr>
<th>Source</th>
<th>Target text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>En</td>
</tr>
<tr>
<td>En</td>
<td>3</td>
</tr>
<tr>
<td>Es</td>
<td>3</td>
</tr>
<tr>
<td>Sl</td>
<td>3</td>
</tr>
</tbody>
</table>

2.3 Selection of lectures

It was agreed the selection of automatic transcriptions to be manually supervised would be made on the basis of transcription quality and topic.

The criterion of transcription quality (low, medium, high), is so that we have an equal distribution of texts for each quality level. If it is not possible for the distribution to be equal among all quality levels due to a constraint caused by the total number of lecture hours to be manually supervised and the duration of the available lectures, preference would be given to lectures with a higher quality level.

The topics of the selected lectures also need to be representative of the entire corpus, while respecting the transcription quality criterion above.

It was agreed that a total of 3 hours of program material per audio language (En, Es, Sl) would be used for manual supervision of transcriptions by language experts in this first round of evaluation. Once transcription files were supervised, these same transcription files would be automatically translated into the target languages involved in trans Lectures. These automatic translations are the ones that will be 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: En↔Sl, En↔Es, En→Fr, En→De) are more than the transcription languages (3 languages in total: En, Es, Sl).

A summary of the amount of material transcribed per language and translated per translation directions is shown in Table 1.

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

3 Transcription Quality

What follows is an analysis of the transcription quality per language of the lectures that were manually supervised. A summary of the editors’ comments is provided. An analysis of the automatic quality metrics for these lectures are also presented.

3.1 Quality control on English transcriptions (DDS)

There were 5 automatic transcription of lectures evaluated in total, 2 of which were of high confidence, 2 medium and 1 low. English automatic transcriptions were generated by EML.

Lecture Id 2430 (low confidence): The ASR output in this instance was completely unusable, and the editor ended up deleted the automatic transcription and transcribing the file
from scratch. The editor described the process as slow, confusing and distracting to try and match the ASR output with the words on the screen. There were no accompanying slides to this lecture although one could tell from the recording that some were used during the lecture. Our editor managed to locate the .pdf file of the lecture in the internet, which proved to be very useful in order to make sense of what would have otherwise been classified as indistinct output. The editor spent more time amending the ASR output in this instance than he would have spent if he had transcribed the file from scratch without resource to the ASR output.

Lecture Id 4408 (medium confidence): The ASR output for this lecture was of low quality. There were no accompanying slides to use and it took a lot of research on the part of our editor to verify the correct spelling of proper names (e.g. of researchers) etc. The accent of the speaker was also occasionally problematic leading to mistakes in the transcription. Finally, the lack of segmentation (with the first half the file being a single large chunk of text) made the file very difficult to work with. Overall it took out editor just as much time to edit the file as it would have taken him to transcribe it from scratch.

Lecture Id 599 (medium confidence): The ASR output in this instance was also of low quality, probably because of the difficult subject matter of this particular lecture that involved a lot of misheard algebra that the editor had to make sense of. The process was very slow, perhaps even slightly slower than originating the file from scratch but definitely no faster. The accompanying slides to the lecture were a very useful resource for the editor who consulted them frequently, but he did not find any evidence of them having been used to provide context to the automatic transcription.

Lecture Id 3369 (high confidence): The ASR output for this lecture was only marginally better than that of Lecture Id 4408 or Lecture Id 599 and it took our editor almost just as long to edit the file than it would have taken him to transcribe it from scratch.

Lecture Id 5678 (high confidence): The ASR output for this lecture was the best one of all five according to our editors. The ASR output was mostly correct, but the chemical formulae and terminology used took a lot of research. Overall, this file was edited in approximately 50% of the time it would have taken an editor to transcribe it from scratch. Our editors classified Lecture Id 5678 as the only file which made their editing work noticeably faster than transcribing from scratch. They attribute this to a combination of factors: the sound quality of the video was good, the speaker was eloquent using good syntax and few hesitations, and the subject matter was not too specialized.

Overall, the quality of the English automatically transcribed lectures was low according to the editors used to review the files. This was evidenced by the fact that overall it took our editors almost just as much time to review and amend the automatic transcriptions than it would have taken them to transcribe them from scratch. An important factor that lowered the bar in terms of transcription quality was the lack of segmentation in the files. Also, the difficulty of the subject matter in many of the lectures probably had a big role to play in the low quality result, as the level of difficulty of the source audio and subject matter was not among the criteria used to select the lectures for manual supervision.

Automatic WER figures are also provided for these lectures in Table 2. As expected, WER figures are correlated with confidence values and human evaluations mentioned above.

3.2 Quality control on Slovenian transcriptions (DDS)

There were 3 automatic transcription of lectures evaluated in total, 1 of which was of high confidence, 1 medium and 1 low. Slovenian automatic transcriptions were generated by EML.

Lecture Id 14881 (low confidence): This lecture was a combination of the speakers actual voice while speaking and his pre-recorded voice-off. The parts where the actual speakers
voice was heard were of good quality overall, but the pre-recorded parts were not recognised properly or not recognised at all in most cases, which resulted in text missing from the automatic transcriptions altogether. The result in those passages was similar to the recognition of background speech/noise. Furthermore, the voice-off parts seemed to block the flow of speech of the directly recorded speech, as the quality of all passages of actual speech that followed voice-off speech were of much worse quality than the rest.

**Lecture Id 8186 (medium confidence):** The quality of this automatic transcription was very low and our editor reported that it took her longer to edit the automatic transcription than it would have taken her to transcribe the file from scratch. It is very possible that the choice of speaker played a very important role in this result. The speaker has a very strong Primorska (South-Western) accent, which makes his vowels a lot wider. He also speaks at a very fast pace, whilst not articulating his words properly. On top of this his grammar is incorrect in the largest part of this talk. A lot of the speech was actually not transcribed at all, whilst very few words were actually transcribed correctly.

**Lecture Id 2409 (high confidence):** The quality of this file was not much better than the other two, although the issues in this case were of a different nature. The recognised file contained large blocks of text without clear delineation of sentence boundaries whilst including a lot of notes regarding hesitations and unrecognized text. Our editor ended up spending a lot of time trying to make sense of the sentences in the recognized text and adding in all the text that was missing in the first place.

On the whole, in the case of Slovene transcriptions, none of the manually supervised files were of good quality. Our editors attributed two main potential reasons for this result. On the one hand, Slovene is a very complex language with approximately 50 different accents, which would naturally cause recognition problems, especially in the case of homonyms, as one word in one accent can sound like another word in another accent. The complexity of the Slovene grammar seemed to make things worse and this was topped by the wide variety of the domains from which the selected lectures were taken.

WER values are presented in Table 3. As happened with English transcriptions, WER values and confidence measures are correlated.

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

<table>
<thead>
<tr>
<th>Lecture Id</th>
<th>Confidence</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>2430</td>
<td>Low</td>
<td>81.5</td>
</tr>
<tr>
<td>4408</td>
<td>Medium</td>
<td>44.0</td>
</tr>
<tr>
<td>599</td>
<td>Medium</td>
<td>43.6</td>
</tr>
<tr>
<td>3369</td>
<td>High</td>
<td>25.5</td>
</tr>
<tr>
<td>5678</td>
<td>High</td>
<td>26.2</td>
</tr>
</tbody>
</table>

Table 3: Summary of WER figures on supervised Slovenian lectures.

<table>
<thead>
<tr>
<th>Lecture Id</th>
<th>Confidence</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>14881</td>
<td>Low</td>
<td>63.0</td>
</tr>
<tr>
<td>8186</td>
<td>Medium</td>
<td>47.9</td>
</tr>
<tr>
<td>2409</td>
<td>High</td>
<td>36.7</td>
</tr>
</tbody>
</table>
3.3 Quality control on Spanish transcriptions (UPVLC)

A total of 3 hours of video lectures in Spanish from the poliMedia repository were selected for the manual revision and evaluation of their automatic transcriptions (generated at EML), based on transcription confidence: 1 hour of high confidence transcriptions (7 lectures), 1 hour of medium confidence transcriptions (10 lectures) and 1 hour of low confidence transcriptions (11 lectures). A table with the details of the 28 selected lectures can be found in Appendix A.4.

Transcription quality was diverse from lecture to lecture. Sound quality was generally not a determining factor in this respect, as all poliMedia video lectures are recorded in a dedicated recording studio, and so audio quality was similarly high for most lectures, even for the ones in the low confidence set.

Transcription quality depended greatly on the speaker and their speaking abilities. Lectures in which the speaker enunciated each word clearly, used short and well-structured sentences and kept a correct rhythm of speech (e.g. Lecture Id 4, Lecture Id 26) had been better transcribed, word by word and sentence by sentence, than lectures in which the speaker spoke in rushed manner, used improvised sentences or tended to mispronounce words (e.g. Lecture Id 23, Lecture Id 10).

The topics covered were different from lecture to lecture, which also had a great impact on transcription quality. The automatic transcription system seemed to be better prepared for some topics than others, and so in some transcriptions most technical terms were correctly transcribed (e.g. Lecture Id 26, on architecture), while in others they were frequently mistranscribed (e.g. Lecture Id 24, on signal processing). The automatic transcriptions of lectures on less technical topics, with few specific technical terms, were less affected by this type of issue (e.g. Lecture Id 5, on the use of an application).

Slides are always clearly visible in poliMedia videos, so they were of great help to the editors when they had to correct any wrongly recognised technical terms.

As to the segmentation of the lectures in discrete sentences, automatic transcriptions came unsegmented. Segmentation was manually performed during the editing/evaluation process. The effort needed for segmentation was lower in the lectures with the more fluent speakers, as they used shorter, clearer sentences and their rhythm marked more distinctly the best points for segmentation.

A few lectures had special characteristics which affected transcription quality and which are worth mentioning. Details on these lectures are presented below:

- Lecture Id 5 (low confidence): Low sound quality. There was a constant background noise in the recording, but the speech was still easily recognisable. Thus, the automatic transcription was full of noise tags, but after removal of the tags we found that the speech had been generally well transcribed (whole sentences were usable with few corrections).

- Lecture Id 9 (low confidence): The speaker spoke Spanish with a foreign (French) accent. Transcription quality was low.

- Lecture Id 18 (medium confidence): The speaker spoke with an American Spanish accent. There were several mistranscriptions due to this fact, especially in terms with /s/ sounds which would have been pronounced /θ/ in standard European Spanish.

- Lecture Id 19 (medium confidence): In this video lecture there were 2 speakers. Their text was completely improvised, with no clear topic, and they interrupted each other. Transcription quality was low.
The total time needed to edit the whole set of 3 hours of automatic transcriptions was approximately 30 hours. The effort required to edit the automatic transcriptions was thus similar, in the current state, to the effort it would have taken to manually transcribe the lectures (taking as reference the manual transcription effort figures recorded in WP2).

In the case of the lectures with the best automatic transcriptions (usually the ones from the speakers showing the best speaking abilities, e.g. Lecture Id 22), the time needed for manual revision was much shorter, and it required less effort than transcribing from scratch.

WER and confidence figures on a representative set of supervised Spanish lectures are provided in Table 4. As observed, WER figures are significantly lower than those of English and Slovenian transcriptions. This fact is explained by an identical recording setup and a large majority of native speakers that made acoustic conditions uniform across lectures.

Table 4: WER and confidence figures on a representative set of supervised Spanish lectures.

<table>
<thead>
<tr>
<th>Lecture Id</th>
<th>Confidence</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Low</td>
<td>43.3</td>
</tr>
<tr>
<td>3</td>
<td>Low</td>
<td>41.9</td>
</tr>
<tr>
<td>9</td>
<td>Low</td>
<td>36.8</td>
</tr>
<tr>
<td>18</td>
<td>Medium</td>
<td>29.7</td>
</tr>
<tr>
<td>5</td>
<td>Low</td>
<td>24.2</td>
</tr>
<tr>
<td>23</td>
<td>High</td>
<td>23.6</td>
</tr>
<tr>
<td>4</td>
<td>Low</td>
<td>21.6</td>
</tr>
<tr>
<td>1</td>
<td>Low</td>
<td>20.9</td>
</tr>
<tr>
<td>25</td>
<td>High</td>
<td>20.5</td>
</tr>
<tr>
<td>12</td>
<td>Medium</td>
<td>19.4</td>
</tr>
<tr>
<td>19</td>
<td>Medium</td>
<td>17.9</td>
</tr>
<tr>
<td>26</td>
<td>High</td>
<td>16.6</td>
</tr>
<tr>
<td>15</td>
<td>Medium</td>
<td>15.3</td>
</tr>
<tr>
<td>28</td>
<td>High</td>
<td>15.2</td>
</tr>
<tr>
<td>24</td>
<td>High</td>
<td>14.9</td>
</tr>
<tr>
<td>27</td>
<td>High</td>
<td>12.8</td>
</tr>
<tr>
<td>22</td>
<td>High</td>
<td>10.3</td>
</tr>
</tbody>
</table>

4 Translation Quality

What follows is an analysis of the translation quality per language pair of the lectures that were manually supervised. A summary of the editors’ comments is provided. An analysis of the automatic quality metrics for these lectures is also provided.

4.1 Quality control on Slovenian into English translations (DDS)

There were 3 files translated from Slovenian into English in total, which were then manually supervised. The automatic translations were provided by RWTH. Overall the quality of the automatic translations was average.

Out of the 3 files automatically translated, Lecture Id 2409 seems to be the better one in terms of quality. However, all lectures still caused many problems to the editor who reported that it took her almost as much time to edit the automatic translations as it would have taken her to translate the files from scratch.
The reason for this was attributed by our editor mainly to the quality of the lectures themselves, rather than the quality of the automatic translations. Both Lecture Id 8186 and Lecture Id 2409 involved speakers that were not constructing their sentences properly in terms of syntax and were jumping from one idea to the next so it was difficult to make sense of what they were talking about even just listening to them in Slovenian. It is only natural that a badly phrased source text would lead to the less than ideal translation. In the case of Lecture Id 14881, which was the shorter lecture, there were no such problems, the lecturer spoke almost as if he was reading his lecture from a script, and used many short sentences, which should have resulted in a translation of a better quality. This did not happen, possibly because of the highly specialised nature of this particular lecture, the topic of which was more difficult from a translation point of view than the other two. The most recurrent mistakes in the automatic translations from Slovene into English were:

- Word order.
- Recognition of nouns when declined or verbs when conjugated.
- Use of present perfect instead of past simple.
- Use of articles before nouns (“the”, “a”, etc.), which were often wrong, missing, or used unnecessarily.

A comparison of manual evaluations and automatic quality scores is provided in Table 5. We can observe that automatic scores are correlated with manual evaluation scores. Moreover, automatic scores computed on post edited translations are higher than those presented in deliverable D6.1.1, in which a single reference translation was available.

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

<table>
<thead>
<tr>
<th>Lecture Id</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Sents.</th>
<th>Avg. Score</th>
<th>BLEU</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>14881</td>
<td>12.7%</td>
<td>21.8%</td>
<td>36.4%</td>
<td>21.8%</td>
<td>7.3%</td>
<td>55</td>
<td>2.9</td>
<td>31.4</td>
<td>49.8</td>
</tr>
<tr>
<td>2409</td>
<td>11.8%</td>
<td>24.7%</td>
<td>32.3%</td>
<td>21.6%</td>
<td>9.7%</td>
<td>672</td>
<td>2.9</td>
<td>30.7</td>
<td>54.6</td>
</tr>
<tr>
<td>8186</td>
<td>4.5%</td>
<td>10.5%</td>
<td>35.4%</td>
<td>37.5%</td>
<td>12.1%</td>
<td>531</td>
<td>3.4</td>
<td>37.8</td>
<td>43.7</td>
</tr>
<tr>
<td>Total</td>
<td>8.7%</td>
<td>18.6%</td>
<td>33.8%</td>
<td>28.3%</td>
<td>10.6%</td>
<td>1258</td>
<td>3.1</td>
<td>33.8</td>
<td>49.7</td>
</tr>
</tbody>
</table>

4.2 Quality control on English into Slovenian translations (DDS)

There were 5 files translated from English into Slovenian in total which were then manually supervised. The automatic translations were provided by RWTH. Overall the quality of the automatic translations was very poor. The main reason for this was failure to follow the rules of the Slovene grammar; verb declinations were not followed and the translation seemed to follow the word order of the source language (English) rather than that of the Slovene language.

Apart from the problems with Slovene grammar, a list of the most common issues encountered is as follows:

1. Some English words are not recognised at all and are left untranslated, including common ones such as “goes”, “a”, “we”, etc.
2. The apostrophe doesn’t seem to be recognised and relevant words are not translated at all, e.g. “I’m”.

10
3. Numbers spelled out in English were frequently left untranslated (e.g. in Lecture Id 4408, but were translated correctly in Lecture Id 5678).

4. Some words that are consistently mistranslated: "organiser" as "potovanje" ("trip, travel"), "speaker" as "predsednik" ("president"), "us" as "ZDA" ("USA" or "the US"), "a few" as "redki" ("few"), "where" as "e" ("if"), "stuff" as "za ivali" ("for animals").

5. Where a word has different possible translations depending on the domain, often the wrong translation was used, e.g. "tell" as "narekovati" ("to dictate"), "interest" as "obresti" ("interest" in the financial sense).

6. Abbreviations were frequently mistranslated with random (incorrect) translations.

7. Slovakian words were used in the translations as well (e.g. mue).

8. Proper names were not capitalised in the translation.

9. The automatic translations did not seem to recognise common phrases (such as "Thanks very much") and mistranslates them, or does not translate them at all ("No" was not translated in one instance).

10. The same word can be translated as 3 or 4 different words, sometimes in the same sentence/context (e.g. in Lecture Id 599).

11. The word "one-dimensional" ("enodimenzionalen") is consistently written in all caps.

Out of 5 files manually supervised, Lecture Id 2430 seems to be the one with the lowest quality in terms of the translations. Most of the automatic translation did not make any sense at all, which resulted in the editor retranslating the file from scratch and assigning a quality score of 1 (lowest) in most cases. Lecture Id 3369 was only marginally better.

Lecture Id 5678 was of low quality also, mainly due to the fact that the lecture was on organic chemistry and a lot of the terms were not translated at all but left in English. It did have a few instances where the editor assigned a 3 score in terms of the quality but these were very few and far between.

Lecture Ids 4408, 3369 and 599 were of slightly better quality overall than the other two files, with a few 2's and 3's allocated in some sentences as well. This was because they included sentences in which there were words that could be edited in order to reach a correct translation (e.g. the root of the word was correct but it was the ending that needed amended in order to be grammatically correct).

On the whole, the experience of editing these pre-translated files proved to be more time consuming that translating from scratch, as the editor had to make sense of the pre-translated text only a very small percentage of which was usable. Short sentences seemed to work better for MT than longer ones, but even those were sometimes completely wrong. If we compare the quality of automatic translations from Slovenian into English and from English into Slovenian, it seems that the former work better, possibly because of the more complex nature of the Slovene grammar. A comparison of manual evaluations and automatic quality scores is provided in Table 6.

4.3 Quality control on English into German translations (DDS)

There were 5 files automatically translated from English into German in total, which were then manually supervised. The automatic translations were provided by RWTH. The quality of the automatic translations is low overall. The main reason for this has to do with the fact that the
Table 6: Summary of manual evaluation metrics vs. automatic quality metrics in English into Slovenian translations.

<table>
<thead>
<tr>
<th>Lecture Id</th>
<th>Sents.</th>
<th>Avg. Score</th>
<th>BLEU</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>2430</td>
<td>93.8%</td>
<td>0.7%</td>
<td>8.5</td>
<td>76.5</td>
</tr>
<tr>
<td>3369</td>
<td>70.3%</td>
<td>2.2%</td>
<td>12.5</td>
<td>72.5</td>
</tr>
<tr>
<td>4408</td>
<td>31.2%</td>
<td>2.1%</td>
<td>12.1</td>
<td>72.5</td>
</tr>
<tr>
<td>5678</td>
<td>59.0%</td>
<td>0.3%</td>
<td>12.5</td>
<td>76.4</td>
</tr>
<tr>
<td>599</td>
<td>81.2%</td>
<td>1.5%</td>
<td>9.7</td>
<td>72.8</td>
</tr>
<tr>
<td>Total</td>
<td>61.3%</td>
<td>1.4%</td>
<td>11.4</td>
<td>72.6</td>
</tr>
</tbody>
</table>

Automatic translations do not seem to take into account German syntax, but rather follow the English syntax of the source files, and hence include sentence constructions that are nonsensical in German.

Interestingly enough, Lecture Id 599 was the file that the editor considered to have the lowest quality in terms of automatic translation into German, with Lecture Ids 2430, 4408 and 3369 only marginally better, while Lecture Id 5678 was considered to be the best out of the 5 files evaluated. While the latter was indeed the best file in terms of quality, as the table above shows, the quality in Lecture Id 599 is certainly not the worse as per the table above. This comment on the part of the editor could be explained because the highly specialised topic of the lecture was exacerbated by the particularly low quality in the speaker’s use of English, which created serious comprehension issues for the editors in all language combinations out of English.

In the case of Lecture Id 599, only short sentences in the file made some sense. The main problem was the syntax, as the MT followed the English syntax, and this was exacerbated by the fact that the source audio involved a lot of hesitations and constructions where the speaker would change syntax in the middle of the sentence, so it was hard to follow him even in English. There were also several mistranslations, especially in terminology. The editor reported it would have taken her less time to translate the file from scratch.

The automatic translation for Lecture Id 2430 was of very low quality. The syntax issues were repeated in this file also, albeit not to the extent that they occurred in Lecture Id 599, and over half of the file contained segments that were nonsensical and had to be re-translated from scratch by entirely discarding the MT output. The subject matter was highly specialised and required a lot of research, even if this was simply to double-check that the MT had translated the terminology accurately, which it did in many instances. Finally all the hyphenated words appeared to be problematic for MT, being left in English most of the time. The editor reported it would have taken her less time if she had translated this file from scratch.

Lecture Id 4408 was also of very low quality, mainly for sentence structure reasons as in the other two files above. The editor reported it would have taken her less time had she translated this file from scratch as well. A possible reason for this result could be the low quality of the source text, as the speaker’s use of English was poor.

Lecture Id 3369 was not better in terms of quality either, as this file had the same syntax and sentence structure issues as all the files above. Also, numbers spelled out in English were frequently left untranslated while abbreviations were frequently mistranslated with random (incorrect) translations. As the lecture was on a highly specialised topic, it also required a lot of research, while most of the terminology was consistently either wrong or left untranslated. The editor reported it would have taken her roughly as much time to do the translation from scratch.

Although Lecture Id 5678 was the best file out of ones manually supervised, there were still so many problems with the translation that even in this case the editor reported it would have
taken her almost just as much time to translate the file from scratch. The main problems lied in the terminology, most of which was consistently mistranslated and with the German word order as in the rest of the files in this language pair. A comparison of manual evaluations and automatic quality scores is provided in Table 7.

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

<table>
<thead>
<tr>
<th>Lecture Id</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Sents.</th>
<th>Avg. Score</th>
<th>BLEU</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>2430</td>
<td>57.2%</td>
<td>18.6%</td>
<td>10.3%</td>
<td>4.8%</td>
<td>9.0%</td>
<td>145</td>
<td>1.9</td>
<td>22.2</td>
<td>57.6</td>
</tr>
<tr>
<td>3369</td>
<td>38.9%</td>
<td>32.4%</td>
<td>17.3%</td>
<td>5.4%</td>
<td>5.9%</td>
<td>185</td>
<td>2.1</td>
<td>19.9</td>
<td>62.3</td>
</tr>
<tr>
<td>4408</td>
<td>50.5%</td>
<td>22.5%</td>
<td>15.3%</td>
<td>5.4%</td>
<td>6.2%</td>
<td>386</td>
<td>1.9</td>
<td>22.9</td>
<td>59.1</td>
</tr>
<tr>
<td>5678</td>
<td>42.3%</td>
<td>37.4%</td>
<td>15.3%</td>
<td>4.3%</td>
<td>0.6%</td>
<td>163</td>
<td>1.8</td>
<td>18.1</td>
<td>66.8</td>
</tr>
<tr>
<td>599</td>
<td>21.3%</td>
<td>46.5%</td>
<td>22.4%</td>
<td>6.3%</td>
<td>3.5%</td>
<td>286</td>
<td>2.2</td>
<td>29.0</td>
<td>52.1</td>
</tr>
<tr>
<td>Total</td>
<td>41.2%</td>
<td>31.6%</td>
<td>16.7%</td>
<td>5.4%</td>
<td>5.1%</td>
<td>1165</td>
<td>2.0</td>
<td>23.3</td>
<td>58.9</td>
</tr>
</tbody>
</table>

4.4 Quality control on English into French translations (DDS)

There were 5 files automatically translated from English into French in total, which were then manually supervised. The automatic translations were provided by XRCE. The quality of the automatic translations was average overall.

Lecture Id 2430 was the file with the lowest quality in terms of automatic translation into French. The MT often did not make any sense, so the editor had to spend a lot of time to decipher it in order to edit it, whilst a lot of research was involved regarding the technical terms used. Some words were translated out of context (e.g. phone), proper nouns were translated in lower case and hyphenated words were almost systematically left in English. Overall, the editor reported that it would have been quicker to translate the file from scratch.

The automatic translations for Lecture Ids 4408, 599 and 5678 were slightly better overall, but all files still required heavy editing and research. Mistranslations (e.g. words translated out of context) as well as grammatical and syntactical errors abound in the files, as the automatic translations do not seem to cater for verb conjugation and noun/adjective declination. Proper nouns were translated in lower case, e.g. grnwald, horstein, bajd, reswick, vodovnik, while hyphenated words are almost always consistently left untranslated (in English), e.g. d-dimensional, d-minus-one-dimensional, zero-mean, carry-over, constraint-induced, multi-channel, six-channel, acid-base, half-equivalence, double-check, which is also the case with various numbers, e.g. sixty-seven, ninety-seven, ninety-eight.

The best file in terms of automatic translation quality was Lecture Id 3369. A lot of editing was still necessary, but the MT output was useful: hardly any 1 or 2 scores were assigned in this file. Mistakes such as the ones above (proper nouns translated in lower case and hyphenated words left in English) also abounded in this file. Our editor classified this file as of above average quality. A comparison of manual evaluations and automatic quality scores is provided in Table 8.

4.5 Quality control on English into Spanish translations (DDS)

There were 5 files automatically translated from English into Spanish in total, which were then manually supervised. The automatic translations were provided by UPVLC. The translation quality of these lectures as reported by our editors was low overall, although the human evaluation metrics indicate that it should be classified as average. When questioned about this, our
Table 8: Summary of manual evaluation metrics vs. automatic quality metrics in English into French translations.

<table>
<thead>
<tr>
<th>Lecture Id</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Sents</th>
<th>Avg. Score</th>
<th>BLEU</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>2430</td>
<td>2.1%</td>
<td>34.5%</td>
<td>37.2%</td>
<td>11.7%</td>
<td>14.5%</td>
<td>145</td>
<td>3.0</td>
<td>40.2</td>
<td>43.6</td>
</tr>
<tr>
<td>3369</td>
<td>0.5%</td>
<td>7.6%</td>
<td>57.8%</td>
<td>23.8%</td>
<td>10.3%</td>
<td>185</td>
<td>3.4</td>
<td>44.5</td>
<td>39.5</td>
</tr>
<tr>
<td>4408</td>
<td>0.3%</td>
<td>5.0%</td>
<td>64.2%</td>
<td>15.1%</td>
<td>15.4%</td>
<td>383</td>
<td>3.4</td>
<td>44.6</td>
<td>39.6</td>
</tr>
<tr>
<td>5678</td>
<td>0.0%</td>
<td>3.0%</td>
<td>60.4%</td>
<td>20.9%</td>
<td>15.7%</td>
<td>134</td>
<td>3.5</td>
<td>51.6</td>
<td>33.0</td>
</tr>
<tr>
<td>599</td>
<td>0.0%</td>
<td>7.7%</td>
<td>57.0%</td>
<td>23.5%</td>
<td>11.8%</td>
<td>272</td>
<td>3.4</td>
<td>44.5</td>
<td>38.1</td>
</tr>
<tr>
<td>Total</td>
<td>0.4%</td>
<td>9.7%</td>
<td>57.5%</td>
<td>18.9%</td>
<td>13.6%</td>
<td>1119</td>
<td>3.4</td>
<td>45.1</td>
<td>38.8</td>
</tr>
</tbody>
</table>

editors responded that, thinking back, they probably awarded higher scores to sentences that were not too difficult to correct, although the text was pretty bad. Also, when the file is riddled with mistakes, editors would be glad to recognize some correct output and might have rewarded this with higher marks than they should have. The human evaluation metrics are provided in Table 9 and the editors’ comments are also summarized below.

The problems encountered in the translations into Spanish were of various types. Terminology was probably the biggest issue, with many terms left untranslated (e.g. Lecture Id 2430) or mistranslated. Strangely enough there were even cases where the same term would be translated wrongly and sometimes not translated at all (e.g. Lecture Id 599). Grammatical and syntactical issues also abounded in the files, mainly to do with word order (in Spanish nouns precede adjectives whereas in English the opposite is the case, and the English word order was commonly used in the files instead of the Spanish one), while verb conjugation was problematic in all files also (e.g. many verbs were not properly inflected but used in the infinitive). Other types of issues encountered are as follows:

- Translation out of context, e.g. the word 'point' is sometimes translated as 'punto' and others as 'coma', when it should always be translated as 'coma' when used in decimal numbers (Lecture Id 5678).
- Acronyms were not recognised as such and frequently translated as words (e.g. Lecture Id 599).
- The use of commas simply followed what was there in the original and many necessary commas were missing (e.g. Lecture Id 599).
- Gender and number concordance issues (e.g. Lecture Id 599, Lecture Id 2430).
- Recurrent mistakes in the translation of numbers (e.g. Lecture Id 2430).
- Wrong accentuation, e.g. the words 'how', 'where', 'when', etc. in Spanish sometimes need an accent and sometimes they don’t, but the MT does not always distinguish properly (Lecture Id 599).
- Semantic pleonasms, e.g. some words don’t need translation in Spanish as they might sound repetitive, but they are still translated (Lecture Id 599).

The editors reported in all cases that it would have taken them the same or less time to have translated the files from scratch. This can be attributed to the highly technical nature of the files which required a lot of research even if only to corroborate that the MT output was indeed correct (which for instance it was at approximately 50% of the time in Lecture Id 5678). The editors also had to spend a lot of time deciphering the MT output in order to edit it. One of the
comments made was that editing such MT output actually results in worse translations quality-wise than if the files had been translated from scratch as the editor is invariably conditioned by the text s/he is provided with, so instead of providing an excellent translation, sometimes the edits only help make the text understandable.

All files bar Lecture Id 599 have been marked with similar scores in terms of quality. The one file that stands out is Lecture Id 599, although this should be taken with a pinch of salt as a different editor worked on this file than the editor that worked on the other four and it is quite possible that she was more lenient in allocating scores as the oral feedback provided on the files as well as they types of errors analysed and detailed in the files themselves seem to be of the same gravity as in the rest of the files. A comparison of manual evaluations and automatic quality scores is provided in Table 9.

Table 9: Summary of manual evaluation metrics vs. automatic quality metrics in English into Spanish translations.

<table>
<thead>
<tr>
<th>Lecture Id</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Sents.</th>
<th>Avg. Score</th>
<th>BLEU</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>2430</td>
<td>48.4%</td>
<td>14.8%</td>
<td>14.8%</td>
<td>7.1%</td>
<td>14.8%</td>
<td>155</td>
<td>2.3</td>
<td>41.2</td>
<td>41.3</td>
</tr>
<tr>
<td>3369</td>
<td>16.3%</td>
<td>29.6%</td>
<td>21.2%</td>
<td>20.2%</td>
<td>12.8%</td>
<td>203</td>
<td>2.8</td>
<td>50.1</td>
<td>31.5</td>
</tr>
<tr>
<td>4408</td>
<td>0.0%</td>
<td>4.8%</td>
<td>34.6%</td>
<td>36.6%</td>
<td>24.0%</td>
<td>413</td>
<td>3.8</td>
<td>35.8</td>
<td>43.2</td>
</tr>
<tr>
<td>5678</td>
<td>18.9%</td>
<td>29.4%</td>
<td>23.1%</td>
<td>18.2%</td>
<td>10.5%</td>
<td>143</td>
<td>2.7</td>
<td>52.0</td>
<td>30.9</td>
</tr>
<tr>
<td>599</td>
<td>15.1%</td>
<td>22.9%</td>
<td>37.3%</td>
<td>16.2%</td>
<td>8.5%</td>
<td>271</td>
<td>2.8</td>
<td>47.3</td>
<td>32.2</td>
</tr>
<tr>
<td>Total</td>
<td>14.9%</td>
<td>17.5%</td>
<td>28.9%</td>
<td>23.0%</td>
<td>15.7%</td>
<td>1185</td>
<td>3.1</td>
<td>43.9</td>
<td>36.8</td>
</tr>
</tbody>
</table>

4.6 Quality control on Spanish into English translations (UPVLC)

A total of 3 hours of automatically translated poliMedia video lectures were manually revised and evaluated. These translations were generated at UPVLC from the revised transcriptions resulting from the manual evaluation process described in Section 3.3. A table with details of the 28 lectures evaluated can be found in Appendix A.4. The human evaluation metrics are provided in Table 10 and the editors’ comments are also summarized below.

Overall, there seemed to be two main factors impacting the quality of the translations: the competence of the speaker when presenting their subject (closely tied to the level of preparation of the lectures) and the topic of the lecture itself.

Lectures in which the speaker used complete grammar structures, not overly colloquial language and short, simple sentences (e.g. Lecture Id 12) resulted in higher-scoring, more intelligible automatic translations than lectures in which the speaker spoke in longer, complex sentences with dubious or incomplete grammar structures, false starts or incorrect word usage (e.g. Lecture Id 23). Sentence length seemed to be a decisive factor: in general, shorter sentences were awarded higher scores.

Meanwhile, reflecting the composition of the poliMedia repository itself, the subject matter of the lectures used for evaluation differed greatly from lecture to lecture. In this evaluation process we observed that the automatic translation system seemed to be better prepared for some topics than for others. For example, in some transcriptions most technical terms were correctly translated (e.g. Lecture Id 12, on mathematics), while in others they were frequently mistranslated or simply left untranslated (e.g. Lecture Id 24, on signal processing). Logically, automatic translations of lectures on less technical topics, with fewer subject-specific terms, were less affected by this issue (e.g. Lecture Id 1, on the use of a software application).

Some frequent mistakes in the automatic translation process might be easily avoided or solved automatically through post-processing:
Both American English and British English spellings are used in the translations (e.g. “analyze” (AE) in Lecture Id 27, “characterises” (BE) in Lecture Id 23).

The indefinite article “a” is not changed to “an” before words beginning with a vowel (e.g. “a illegitimate” in Lecture Id 27).

Contractions are used at random (e.g. “we’re” and “we are” in Lecture Id 27).

The adjective plus noun order is frequently reversed (e.g. “a programme infected”, “has code illegitimate” in Lecture Id 27).

The first person pronoun “I” is written in lower-case (e.g. in Lecture Id 18).

“That is to say” is frequently used to translate “es decir” (e.g. in Lecture Id 27), when it should appear more frequently in its simpler form, “that is”.

As a general note, we can say that translation quality in the automatic translation left ample room for improvement. The average quality score across all evaluated translation was 2.86 (from 1 to 5). In more detail, 42% of the segments were given a quality score of 1 or 2, 24% were given a 3, and 34% were given a 4 or a 5.

In terms of the time taken to revise the automatic translations in relation to the time it would have taken to translate them manually, this varied greatly from lecture to lecture. In some cases (e.g. Lecture Id 23) there were no improvements in this area and, arguably, editing/revision may have even taken longer than if the text had been translated from scratch. This is because the translator has to first read, try to understand, piece together and identify the key errors in the automatically-generated segment, before superimposing their own version (or discarding the automatic version entirely).

On average, correcting the automatic translations took a similar time as it would have taken to translate them manually, taking as reference the manual translation effort rates recorded in WP2 (around 40 times the duration of the video). However, in the case of the more highly-scored translations, revision took less time than that estimated for manual translation. A comparison of manual evaluations and automatic quality scores is provided in Table 10.

5 Conclusions

The results of the manual supervision of the automatically transcribed and translated lectures showed that the overall quality of the files was below average/low.

In the case of transcriptions, an important factor that lowered the bar in terms of transcription quality was the lack of segmentation in the files, which was worsened by the complexity of the Slovene grammar in the case of Slovenian. Also, the difficulty of the subject matter in many of the lectures probably had a big role to play in the low quality result, as the level of difficulty of the source audio and subject matter was not among the criteria used to select the lectures for manual supervision. Perhaps this should be a factor to be considered in the second round of manually supervised evaluations. In the case of Slovenian in particular, the fact that there are approximately 50 different accents used in the country also seemed to impact on the quality of automatic transcriptions.

In the case of translations, the general conclusion that can be drawn is that MT performs better when translating into English MT than in the reverse direction, presumably because of the less morphologically complex nature of the English language as well as the larger language models available for English. The quality of the audio in terms of how well constructed the sentences in the source text were made a marked difference in the quality of the machine
Table 10: Summary of manual evaluation metrics vs. automatic quality metrics on a representative set of Spanish into English translations.

<table>
<thead>
<tr>
<th>Lecture Id</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Sents</th>
<th>Avg. Score</th>
<th>BLEU</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>69.8%</td>
<td>20.9%</td>
<td>4.7%</td>
<td>0.0%</td>
<td>4.7%</td>
<td>43</td>
<td>1.5</td>
<td>36.7</td>
<td>45.7</td>
</tr>
<tr>
<td>4</td>
<td>54.6%</td>
<td>27.3%</td>
<td>13.6%</td>
<td>0.0%</td>
<td>4.6%</td>
<td>22</td>
<td>1.7</td>
<td>37.5</td>
<td>39.7</td>
</tr>
<tr>
<td>19</td>
<td>23.1%</td>
<td>61.6%</td>
<td>15.4%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>13</td>
<td>1.9</td>
<td>42.4</td>
<td>37.7</td>
</tr>
<tr>
<td>27</td>
<td>42.0%</td>
<td>29.6%</td>
<td>14.8%</td>
<td>9.9%</td>
<td>3.7%</td>
<td>81</td>
<td>2.0</td>
<td>41.7</td>
<td>41.4</td>
</tr>
<tr>
<td>25</td>
<td>34.8%</td>
<td>34.8%</td>
<td>13.0%</td>
<td>8.7%</td>
<td>8.7%</td>
<td>23</td>
<td>2.2</td>
<td>46.6</td>
<td>39.8</td>
</tr>
<tr>
<td>28</td>
<td>32.0%</td>
<td>32.0%</td>
<td>20.0%</td>
<td>4.0%</td>
<td>12.0%</td>
<td>25</td>
<td>2.3</td>
<td>35.9</td>
<td>46.3</td>
</tr>
<tr>
<td>10</td>
<td>32.4%</td>
<td>29.7%</td>
<td>16.2%</td>
<td>5.4%</td>
<td>16.2%</td>
<td>37</td>
<td>2.4</td>
<td>36.9</td>
<td>42.8</td>
</tr>
<tr>
<td>24</td>
<td>31.3%</td>
<td>25.4%</td>
<td>25.4%</td>
<td>11.9%</td>
<td>6.0%</td>
<td>67</td>
<td>2.4</td>
<td>39.4</td>
<td>43.0</td>
</tr>
<tr>
<td>22</td>
<td>21.7%</td>
<td>13.0%</td>
<td>21.7%</td>
<td>18.8%</td>
<td>24.6%</td>
<td>69</td>
<td>3.1</td>
<td>48.0</td>
<td>33.8</td>
</tr>
<tr>
<td>5</td>
<td>6.3%</td>
<td>18.8%</td>
<td>37.5%</td>
<td>31.3%</td>
<td>6.3%</td>
<td>16</td>
<td>3.1</td>
<td>30.2</td>
<td>42.4</td>
</tr>
<tr>
<td>26</td>
<td>7.5%</td>
<td>17.0%</td>
<td>39.6%</td>
<td>24.5%</td>
<td>11.3%</td>
<td>53</td>
<td>3.2</td>
<td>41.5</td>
<td>41.3</td>
</tr>
<tr>
<td>18</td>
<td>18.2%</td>
<td>3.6%</td>
<td>30.9%</td>
<td>25.5%</td>
<td>21.8%</td>
<td>55</td>
<td>3.3</td>
<td>33.6</td>
<td>43.4</td>
</tr>
<tr>
<td>3</td>
<td>0.0%</td>
<td>18.4%</td>
<td>42.9%</td>
<td>28.6%</td>
<td>10.2%</td>
<td>49</td>
<td>3.3</td>
<td>50.5</td>
<td>30.1</td>
</tr>
<tr>
<td>15</td>
<td>0.0%</td>
<td>11.1%</td>
<td>22.2%</td>
<td>33.3%</td>
<td>33.3%</td>
<td>9</td>
<td>3.9</td>
<td>54.3</td>
<td>31.6</td>
</tr>
<tr>
<td>12</td>
<td>1.1%</td>
<td>6.6%</td>
<td>29.7%</td>
<td>19.8%</td>
<td>42.9%</td>
<td>91</td>
<td>4.0</td>
<td>48.0</td>
<td>33.2</td>
</tr>
<tr>
<td>1</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>54.5%</td>
<td>45.5%</td>
<td>11</td>
<td>4.5</td>
<td>56.4</td>
<td>27.4</td>
</tr>
<tr>
<td>Total</td>
<td>24.1%</td>
<td>20.5%</td>
<td>24.0%</td>
<td>15.8%</td>
<td>15.6%</td>
<td>704</td>
<td>2.8</td>
<td>42.2</td>
<td>39.8</td>
</tr>
</tbody>
</table>

translation. The specialisation of the subject matter on the other hand had the opposite effect, with mistranslations increasing the more specialised the topic of the lecture. As this factor was not considered when selecting the lectures to be manually supervised in this first round of evaluation, perhaps it should be taken into account in the next round. Overall automatic translation quality in all language combinations out of English was judged as low by the editors, whereas in the case of translations into English the quality level rose to average.

For the next quality control, automatic transcriptions are expected to be significantly improved taking advantage of the slide textual content so as to minimise editor supervision effort. Editors’ comments on automatic translations will be very useful to enforce among others, grammar constraints and correct errors that could be fixed by postprocessing automatic translations.
References

A Appendix

A.1 Transcription guidelines (DDS)

1. The transcriber tool (Transcriber 1.5.1) is used for all transcriptions.

- All transcriptions will be UTF-8 encoded.
- For all segments begin and end time markers are set. Please use speaker turn labels to indicate segment ends. To turn on the “Speaker turn label” function, click “Segmentation”, then “Create turn...”

2. Always transcribe what was said.

- Transcribe orthographically correct.
- Do not correct grammar.
- Do not capitalize at the beginning of a sentence.
- Separate punctuation by a white space from preceding word.
- If a punctuation mark is spoken, it will be transcribed as word and not as punctuation mark.

Examples:
- “gonna” simply transcribe “gonna” and not “/gonna/going to/”
- Wrong acronyms (e.g. PCA when PCD should be used) transcribe what was said, e.g. “PCA PCD space”, not “/PCA/PCD/PCD space”
- When speaker corrects himself: transcribe what was said, e.g. “visualisations is are specifically” not “visualisations /is/are/ specifically”.

3. When possible, each segment must represent a complete meaningful sentence.

Example:
Correct: this is a complete, meaningful sentence.
Incorrect: this is a complete,
meaningful sentence.

4. Letters and spelled out words: each letter will be individually typed uppercased, separated from the other letter by a space. Example: my name is Smith S M I T H

For acronyms (e.g.: USA, NSA), the individual letters will not be separated by spaces.

5. Numbers, digits, ordinals, variable names, equations:

- Numbers, digits and ordinals will be transcribed as words.
  Examples: four hundred eighteen, forty-fourth, two thousand and two
- Decimals are not written numerically.
  Example: two point six
- Variables are transcribed as word.
  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, when pronounced oh, must be transcribed with oh.
  Example: X509 must be transcribed /X five oh nine / X509 /
• Proprietary names must be spelled out and // notation must be used.
  Example 1: Web 2.0 must be transcribed /Web two dot oh/web 2.0
  Example 2: three squared plus two, seven point three minus three point eight

6. All transcriptions must have, at least, one speaker. Speakers can be created by clicking the
   initial tag “(no speaker)” at the beginning of the transcription. The following information
   must be defined if available:
   • Speaker first and last name.
   • Gender (male/female).
   • Native/non-native. If non-native: accent (if possible). For Slovene audio files, use
     Slovene instead of Native.
   • If there are several speakers, it is mandatory to use a new turn whenever there is a
     speaker change. A new turn can be created from menu or pressing Ctrl+t.

7. Incomprehensible passages will be marked with the respective transcriber feature. The
   following non-speech events are transcribed with respective transcriber features:
   • Hesitations (voiced sounds). Please note that only non-words should be marked as
     hesitations. Both British and American spelling is fine as long as there is consistency
     and all instances are marked as hesitations.
   • Speaker noise (unvoiced sounds): lip smack, etc. Exception: do not annotate breaths.
     - Non-stationary noise: (door slam, window,...)
   • Music
   • Chatter and noise: Use Ctrl+b for background noise and annotate it as speech and
     noise. This appears in Transcriber as ((o-))((-o))

8. Long silences (more than a second) in the middle of or between sentences are treated as
   follows:
   Insert a new segment with the tag: ~LONGSIL
   Note: if there is a long silence half way through a sentence, use ~LONGSIL within a
   segment. Create a new segment only when the silence falls between sentences.

9. Long pauses: Click “Segmentation”, then “Create turn...” (type = nontrans) and when
   silence is over start a new (transcribed) section.

10. Speech disfluencies: for such language phenomena we will use the // notation, that can
    be summarized as: /what the speaker says/what the speaker wanted to say/. It is easier
    to understand when seeing some examples:
    • Incorrect pronunciation: the speaker utters something incorrectly. Incorrect pronun-
      ciations are only transcribed for native speakers!
      Example:
      Speaker says: ... there was a 'prublem' with ...
      Transcription: ... there was a /prublem/problem/ with ...
      Explanation: The speaker pronounce ‘prublem’ but he wanted to say ‘problem’
    • Foreign words: for words, titles, etc. in foreign languages that are pronounced dif-
      ferent from English (or Slovenian respectively) pronunciation rules, the same // notation
      will be used.
      Example:
      Speaker says: ... there was a coup d’tat in ...
      Transcription: ... there was a /koo day tah/coup d’tat/...
• Repetition: the speaker, unintentionally, repeats a word or expression.
  Example:
  Speaker says: ... there was a a a problem with ...
  Transcription: ... there was /a// /a// problem with ...
  Explanation: The second and third “a” were uttered but should not appear in the transcription.
  Or we transcribe: ... there was /a a a/a/ problem with ...
  Explanation: The speaker said three “a” but he just wanted to say one.

• Do not label as repetition if its interrupted by hesitation or cut off.
  Example:
  performed every day b∼ by [hes-er][-hes] by the employees the s∼ the same goes for software testing

11. Word cut-offs: are marked by a ∼.
  Example:
  Speaker says: ... can we meet on Thurs Friday
  Transcription: ... can we meet on Thurs ∼ Friday ...
  Note: ∼ should only be attached to words that are cut-off and not when the audio cuts in halfway through a sentence.
  Do not start a new segment after a “sentence cut-off”.

12. Discourse markup: Do not use discourse markup.


14. Quotation marks (reported speech): Avoid quote marks if possible, but where unavoid-
   able, transcribe quotes like other punctuation marks (first letter in lowercase) with spaces between the quote and the quotation marks, for example: “you should do it like this .”.

15. Foreign Dialogue: leave foreign dialogue untranscribed (but labeled) in a separate section entitled nontrans.

16. Overlapping speech: Annotate using the Transcribers overlapping speech function (two speakers will appear in the same segment).

17. Slovene lectures: Use correct grammatical spelling for colloquial terms.
A.2 Translation guidelines (DDS)

1. The transcriber tool (Transcriber 1.5.1) is used for all translations and the final files are in the .trs format.

2. Orthographical standards:
   Please use correct orthographical standards and normal capitalisation conventions per language in all languages.

3. Annotations:
   a) Cases of annotations that appear within square brackets, with speech elements in between or not (e.g. hesitations [hes-]erm[-hes], coughs [toux en fond]), etc.) → These can automatically be removed in the transcribed files and they will not be reproduced in the translated files.
   b) Cases of repetitions followed by silences, where the ~ mark is used. → In such cases repetitions and cut-offs will be translated, so as not to cause any problems with the alignment.
      Example:
      “but if it is a hu~ it is a human, how do you know it’s human?”
      “mais si c’est un hu~ si c’est un humain, comment est-ce qu’on sait que c’est un humain ?”.
   c) LONGSILs → Do not include them in the target text as these need to be removed from the source text before word alignment.

Numbers, digits, etc: Follow the style in the transcriptions (i.e. full words if it is in full words in the source text, numbers if it is in numbers in the source text) for consistency.
A.3 Translation Quality Assessment Guidelines (XRCE)

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

- 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:**
  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
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 estuviera dividido por el VIH protease.

**Spanish human post-edited:**
Los corpúsculos virales infeccioos maduros no se producirían si el virus VIH no fuese 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, diluyendo la proporción debido a los accionistas actuales.

**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 (EML and UPVLC)

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

<table>
<thead>
<tr>
<th>ID</th>
<th>Category</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>2430</td>
<td>Speech Analysis</td>
<td>00:20:21</td>
</tr>
<tr>
<td>4408</td>
<td>Sensor Networks</td>
<td>00:27:35</td>
</tr>
<tr>
<td>599</td>
<td>Machine Learning</td>
<td>00:59:51</td>
</tr>
<tr>
<td>3369</td>
<td>Natural Language Processing</td>
<td>00:25:29</td>
</tr>
<tr>
<td>5678</td>
<td>Chemistry</td>
<td>00:50:04</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>03:03:20</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>ID</th>
<th>Category</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>14881</td>
<td>Physics</td>
<td>00:05:47</td>
</tr>
<tr>
<td>8186</td>
<td>Business</td>
<td>01:18:44</td>
</tr>
<tr>
<td>2409</td>
<td>Environment</td>
<td>01:40:58</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>03:05:29</td>
</tr>
</tbody>
</table>
Table 13: Spanish lectures from poliMedia selected for quality control.

<table>
<thead>
<tr>
<th>ID</th>
<th>Category</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>University administration</td>
<td>00:01:16</td>
</tr>
<tr>
<td>2</td>
<td>University administration</td>
<td>00:04:58</td>
</tr>
<tr>
<td>3</td>
<td>Industrial computer science</td>
<td>00:05:05</td>
</tr>
<tr>
<td>4</td>
<td>Finance</td>
<td>00:04:05</td>
</tr>
<tr>
<td>5</td>
<td>University administration</td>
<td>00:02:04</td>
</tr>
<tr>
<td>6</td>
<td>Intranet UPV</td>
<td>00:00:56</td>
</tr>
<tr>
<td>7</td>
<td>Descriptive statistics</td>
<td>00:08:42</td>
</tr>
<tr>
<td>8</td>
<td>Internet</td>
<td>00:11:12</td>
</tr>
<tr>
<td>9</td>
<td>Logics</td>
<td>00:09:45</td>
</tr>
<tr>
<td>10</td>
<td>Data protection</td>
<td>00:05:18</td>
</tr>
<tr>
<td>11</td>
<td>Intranet UPV</td>
<td>00:06:31</td>
</tr>
<tr>
<td>12</td>
<td>Autocad</td>
<td>00:06:55</td>
</tr>
<tr>
<td>13</td>
<td>Sound Technology</td>
<td>00:07:10</td>
</tr>
<tr>
<td>14</td>
<td>Backbones</td>
<td>00:11:21</td>
</tr>
<tr>
<td>15</td>
<td>Cooperation</td>
<td>00:01:10</td>
</tr>
<tr>
<td>16</td>
<td>Valuation</td>
<td>00:04:15</td>
</tr>
<tr>
<td>17</td>
<td>Education</td>
<td>00:11:35</td>
</tr>
<tr>
<td>18</td>
<td>Education</td>
<td>00:05:36</td>
</tr>
<tr>
<td>19</td>
<td>Institutional relations</td>
<td>00:01:42</td>
</tr>
<tr>
<td>20</td>
<td>Marketing</td>
<td>00:01:30</td>
</tr>
<tr>
<td>21</td>
<td>Agricultural engineering</td>
<td>00:07:12</td>
</tr>
<tr>
<td>22</td>
<td>Sound Technology</td>
<td>00:08:36</td>
</tr>
<tr>
<td>23</td>
<td>Law</td>
<td>00:08:42</td>
</tr>
<tr>
<td>24</td>
<td>Signal processing</td>
<td>00:10:51</td>
</tr>
<tr>
<td>25</td>
<td>Tourism</td>
<td>00:04:52</td>
</tr>
<tr>
<td>26</td>
<td>Construction</td>
<td>00:08:44</td>
</tr>
<tr>
<td>27</td>
<td>Computer science</td>
<td>00:12:44</td>
</tr>
<tr>
<td>28</td>
<td>Operations management</td>
<td>00:04:46</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>03:04:11</td>
</tr>
</tbody>
</table>
## B Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR</td>
<td>Automatic Speech Recognition</td>
</tr>
<tr>
<td>BLEU</td>
<td>Bilingual Evaluation Understudy</td>
</tr>
<tr>
<td>DDS</td>
<td>Deluxe Digital Studios Limited</td>
</tr>
<tr>
<td>EML</td>
<td>European Media Laboratory GmbH</td>
</tr>
<tr>
<td>JSI</td>
<td>Josef Stefan Institute</td>
</tr>
<tr>
<td>K4A</td>
<td>Knowledge for All Foundation</td>
</tr>
<tr>
<td>MT</td>
<td>Machine Translation</td>
</tr>
<tr>
<td>RWTH</td>
<td>RWTH Aachen University</td>
</tr>
<tr>
<td>TER</td>
<td>Translation Error Rate</td>
</tr>
<tr>
<td>UPVLC</td>
<td>Universitat Politècnica de València</td>
</tr>
<tr>
<td>WER</td>
<td>Word Error Rate</td>
</tr>
<tr>
<td>XRCE</td>
<td>XEROX Research Center Europe</td>
</tr>
</tbody>
</table>