This document is part of the Research and Innovation Action "European Live Translator (ELITR)".
This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 825460.

Deliverable D6.6

Final Report on Integration

Chiara Canton (PV), Rishu Kumar (CUNI), Ondřej Bojar (CUNI),
Armin Schweinfurth (AV), Franz C. Krüger (AV)

Dissemination Level: Public

Final (Version 1.0), 31st March, 2022

Grant agreement no.	825460
Project acronym	ELITR
Project full title	European Live Translator
Type of action	Research and Innovation Action
Coordinator	doc. RNDr. Ondřej Bojar, PhD. (CUNI)
Start date, duration	1st January, 2019, 39 months
Dissemination level	Public
Contractual date of delivery	31st March, 2022
Actual date of delivery	31st March, 2022
Deliverable number	D6.6
Deliverable title	Final Report on Integration
Type	Report
Status and version	Final (Version 1.0)
Number of pages	8
Contributing partners	AV, PV, CUNI
WP leader	PV
Author(s)	Chiara Canton (PV), Rishu Kumar (CUNI), Ondřej Bojar (CUNI), Armin Schweinfurth (AV), Franz C. Krüger (AV)
EC project officer	Luis Eduardo Martinez Lafuente
The partners in ELITR are:	▪ Univerzita Karlova (CUNI), Czech Republic ▪ University of Edinburgh (UEDIN), United Kingdom ▪ Karlsruher Institut für Technologie (KIT), Germany ▪ PerVoice SPA (PV), Italy ▪ alfatraining Bildungszentrum GmbH (AV), Germany
Partially-participating party	▪ Nejvyšší kontrolní úřad (SAO), Czech Republic

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

doc. RNDr. Ondřej Bojar, PhD., ÚFAL MFF UK
Malostranské náměstí 25
118 00 Praha, Czech Republic

bojar@ufal.mff.cuni.cz
Phone: +420 951 554 276
Fax: +420 257 223 293

Copies of reports and other material can also be accessed via the project's homepage:

`http://www.elitr.eu/`

Contents

1 Executive Summary

This deliverable reports on the integration of the services and systems developed by ELITR partners into production processes to provide end-to-end workflows for the use cases of the project. In this report the last updates on continuous evaluation (Section 2), efforts to ensure the replicability and robustness of the system (Section 4), last updates on the Minuting Demonstrator (Section 3) not already reported in deliverable "D6.5 Demonstrator of Automatic Minuting" are presented, and on Events (Section 5) not already reported in "D6.4 Report on Follow-up SAO Events".

WP6 Integration is divided into four tasks with the following progress:

T6.1 Platform Updates (months 1-24): The existing PV and AV platform has been updated to be able to handle the new use cases, primarily the multi-target translation and display of subtitles. As a part of this task, a web-based "Publishing platform", the Presentation Platform has been developed, to deliver the translations into the many target languages to the end users. Also the Online Text Flow Application as been developed by CUNI. The activity is completed. For a more complete overview of presentation methods and reasoning on user experience put in place by the ELITR project, please refer to "D6.1 Publishing Platform".

T6.2 Integration of ASR, SLT and multi-target SMT engines (months 3-36): All the components developed by the research partners have been integrated into the PV Platform and have been used both at SAO events and in alfaview® conferencing platform. For a more complete overview on the integration of the partners, please refer to "D6.3 Intermediate Report on Integration".

T6.3 Minuting Demonstrator (months 1-36): The design and implementation of a demonstrator platform that provides live transcript and minuting for participants of online meetings has been completed. It has been integrated with the ASR systems through the alfaview® conferencing platform and the Pervoice Service architecture, see the dedicated Deliverable "D6.5 Demonstrator of Automatic Minuting" and Section 3.

T6.4 Running SAO Events (months 7-36): Starting from the early beginning of the project all the interpreting and subtitling systems have been run at the various live events organized by SAO. Last year took place the EUROSAI Congress, project's main event. See the dedicated deliverables "D6.2 Report on ELITR at EUROSAI Congress" and "D6.4 Report on Follow-up SAO Events" and Section 5 for a brief report of the last events.

This report is concluded in Section 6.

2 Continuous Evaluation

elitr-testset-evaluation serves as our internal pipeline to continuously test the outputs of our ASR and MT system against the reference documents in elitr-testset. This git repository and tools was intended to record the increase/decrease in the performance of our systems and to make the tests reproducible in a deterministic way. The system was introduced in Deliverable D6.3 (Intermediate report on Integration), and since then we have continued to add data to the test set thanks to the continuous work of human translators and annotators. Since the last report, various data has been added and categorized going to enrich old indices and creating new ones. This work allowed us to be able to more accurately choose the best system to use based on the use case. It follows a brief list of activities performed since the last update on this topic, mainly consisting of polishing and updating our index files for elitr-testset, which serves as test-suites for evaluation.

The category of audio spoken by non-native speakers has been enriched with over 40 English audio files, translated manually into Czech and German. The Non-native English speakers are from Spain, Poland, Austria, Germany, Holland, Belgium and Romania.

Several verbatim transcriptions were added for the European Parliament, Czech audio files, and of the data collected during the WG VAT meeting (2021), as well as the broadcasts of the radio station Český rozhlas and of the Chamber of Deputies of the Czech Republic, resulting in a total of 10 hours of verbatim transcribed Czech audio material.

We added also verbatim transcriptions in English of the first LangTools Workshop, a workshop organized by UFAL colleagues for the EUROSAI Congress held in 2020 together witch Czech and Slovak manual translations.

The WMT18 dataset has been added to the *elitr-testset*. It consists of Czech articles which were read sentence by sentence in individual recordings. The sentence-level recordings were then concatenated in a loss-less way in the original OGG format. The same articles were read by Czech native speaker in Czech and in English (where the speaker is non-native). This dataset can thus serve for SLT evaluation in both direction: English or Czech source to the other language.

Also we continued the activities on the Theaitre and related interviews at Robothon 2021, debates held in Czech. We collected both the recording and the live respeaking of the interpreter. The purpose of this testset is to measure the quality of the transcription comparing different inputs: live automatic transcription and reaspeaking automatic transcription.

Particular attention was also dedicated to the revision of verbatim transcripts already produced, through an activity of revision of the capitalization of terms (for the Czech language) and of the split of lines into sentences in order to facilitate the evaluation of automatic machine translation systems.

In order to see the evaluation results, please have a look at deliverable D6.3 (Intermediate report on Integration).

3 Minuting Demonstrator

In order to improve the integration and usability of the Minuting Demonstrator, periodic meetings have been planned between the teams involved (AV, PV, KIT). The meetings were held directly on the alfaview platform in order to produce the automatic transcripts underlying the tests of the automatic minute system. The meeting topics were related to: solve problems encountered in the use of the MCloud library, analyze the quality of the results. The quality of the results has been human-evaluated on the following point of views:

- text output in the Alfaview platform
- quality of the transcription provided by the ASR
- quality/usability of the minute produced automatically

In the rest of this section (Section 3.1), we report on some of the most meaningful observations and debug sessions on the Alfaview platform.

3.1 Debugging/Integration

While integrating the latest changes, we noticed that some transcriptions words were missing or cut off.

1. After some research we noticed that the library was crashing with just the error message "access violation". As we had no access to source code or documentation of the binary library which integrates alfaview with the PerVoice platform, debugging was quite difficult. We went ahead and created a core dump of the crashing application for closer inspection. In multiple joint debugging sessions we were able look into the actual memory state at the time of the crash. Ultimately we isolated a nil pointer exception as the root cause which was then fixed in the process.

2. Additionally it was noticed that at the beginning of transcriptions, sometimes there were inaccuracies or even parts missing entirely. As the behaviour could not be replicated consistently, we figured it had something to do with transmission or connection to the remote ASR workers. After further debugging we noticed that the transcription buffer was only set to 2 seconds in size. Apparently, establishing the connection to the ASR worker often took more than that, which lead data loss and subsequent inaccurate or missing parts of the transcripts. Increasing the buffer size to 20 seconds fixed the problem.

3. When dealing with longer continuous transcriptions, we became aware that at some point the transcription was simply cut off or overwritten by intermediate updates. Again this is not a consistent behaviour and we have not yet found the root cause. We are currently trying to figure out where the problem originates from, alfaview, ASR worker or the mediator. At the time of writing, debugging the issue is still ongoing.

4 Replicability and robustness

We have pursued the need to have the services distributed and replicated also among the partners in order to avoid possible bandwidth problems or unreachable services. For this reason the UEDIN automatic translation service was made available in a dockered version, while the PerVoice orchestration service (mediator) was provided as a normal software package. Both have been distributed to the CUNI partner in order to let them manage possible "offline events" or bandwidth problems.

5 Events

In this section we are reporting events where we provided transcription and translation which are not covered in Deliverable 6.4. As described in the Deliverable 6.4, we offer our transcription and translation service as a research demonstration.

5.1 JSIM

The Day of Informatics and Mathematics at Department of Mathematics and Physics was the first event where we provided our services. This demonstration did not go according to plan and the timestamps of the ASR output kept leaking in the output platform. This unexpected event also resulted in our Czech ASR worker and MT workers crashing constantly without any reasonable bug. On a preliminary analysis, the main issue of the crashes looks like the overloaded system which was used to run the docker image of our pipelines as well as obtaining sound. Further stress testing on our dedicated Virtual Machine servers failed to replicate this behaviour.

5.2 ELG Workshop Serbia

The European Language Grid workshop organised by the University of Belgrade in Serbia, served as the first major event (after the events mentioned in D6.4) where our system worked as intended and we did not experience any major crashes. In the initial hour of the events, we experienced some hiccups, with Zoom webpage not providing us the option to join the meeting from a browser.

The preparation for the event was also done in the same manner as the previous ELG workshops. The organizers provided us with some of the slides used during the presentation beforehand and we used tools such as *pdftotext* to extract keyword from the data which were not present or very scarcely present in the dictionary of our ASR worker for English. These words were added to the *memory* which in turn was used by KIT's ASR system.

During the event, an operator was watching the produced output and kept actively adding words which were new to the ASR worker or had an erroneous transcription. In some instances, it involved human judgement on the likelihood of that word reappearing during the event. This process also includes adding words to the memory which are known during the main training but with a populated memory get confused with some memory words and thus constitute false positives.

This process of insertion of new words is version-controlled with basic Linux operations, so that a retrospect analysis can be carried out on the full history of memory states. A custom *.vimrc* file is used in the directory of the memory file, which on every save, i.e., the `:w` command in the Vim editor runs also the bash oneliner "`git add && git commit -m "$timestamp"`" where timestamp is automatically generated when `:w` command is issued.

6 Conclusion

The key to a successful component integration are communication and collaboration. Thanks to the multidisciplinary team that the project was able to set up, the main integrations were already completed in the first part of the project. The driving force towards integration was the commitment, from the early beginning, to participate in events and demos. Great efforts have been spent to make the system more and more reliable and provide an increasingly robust service, also aiming at measuring the quality of the technologies.

Finally, in the last part of the project, in terms of integration, we invested in the creation of the Minuting Demonstrator. The technology underlying the demo constitutes the most ambitious research objective of the project, and now it can be demonstrated as a service prototype integrated with the rest of the ELITR infrastructure.