Research and Innovation Action (RIA) H2020 – 957017

SELMA

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu/

D4.4 Final platform release with full continuous massive stream learning capabilities

Work Package	4
Responsible Partner	IMCS
Author(s)	Guntis Barzdins, Sebastião Miranda, Didzis Gosko
Contributors	Afonso Mendes, Arturs Znotins, Mikus Grasmanis, Paulis Barzdins, Roberts Dargis, Normunds Gruzitis
Reviewer	Yannick Esteve
Version	1.0
Contractual Date	31 March 2024
Delivery Date	28 March 2024
Dissemination Level	Public

Version History

Version	Date	Description
0.1	16/02/2024	Initial Table of Contents (ToC)
0.2	21/02/2024	Overall deliverable structure drafted
0.3	22/02/2024	Ready for internal review
0.4	14/03/2024	Internal review feedback received, final updates
1.0	14/03/2024	Publishable version

Executive Summary

The final platform release with full continuous massive stream learning capabilities covers two main Use Cases UC1, UC2 of the SELMA project, along with internal testing Use Case UC0. UC0 served as a playground for various NLP technologies developed and trialed during the SELMA project. The most successful of these technologies ended up in our commercial UC1 and UC2 products described in deliverables D1.4 plus D4.5 and D4.6 respectively, as well as in the Podcast Creation use case described in D1.4.

This document focuses on the “privacy”, “scalability” and “open-source” aspects of SELMA project software and transitioning from DockerSpaces cloud approach described in deliverable D4.3 to the private edition SELMA UC0 Open-Source Software as the final solution due to both technical and ethical considerations.

Table of Contents

Executive Summary 3

1. *Introduction* 6
2. *"White" Version of Use Case 0: SELMA NLP Service Orchestration* 8
3. *"Yellow" Version of Use Case 0: FrontEnd NLP Pipeline and DockerSpaces Backend* .. 11
4. *"Green" Version of Use Case 0: SELMA Open-Source Candidate* 14
5. *"Red" Version of Use Case 0: Final SELMA Open-Source Software* 16
6. *Conclusion* 19

Table of Figures

FIGURE 1 *"WHITE" VERSION OF THE SELMA BASIC TESTING AND CONFIGURATION INTERFACE (UC0)* *8*

FIGURE 2 *"YELLOW" VERSION OF THE UC0: FRONTEND NLP PIPELINE WITH UNIVERSAL DOCKERSPACES BACKEND* *11*

FIGURE 3 *"GREEN" VERSION OF THE UC0: SELMA OSS CANDIDATE WITH SCALABLE DOCKERSPACES BACKEND* *14*

FIGURE 4 *"RED" VERSION OF THE UC0: FINAL SELMA OSS WITH STATICALLY LINKED BACKEND FOR APPLE MACOS (ARM M1, M2, M3 OR INTEL) OR LINUX (X86-64) END-USER COMPUTER* *17*

1. Introduction

The final SELMA platform release with *full continuous massive stream learning* capabilities is described in detail in the deliverables D1.4, D4.5 and D.4.6 covering the primary media monitoring Use Case 1 (Monitio) and the primary media production Use Case 2 (plain X). These primary use cases rely on the technologies developed and trialed in the testing Use Case 0 introduced in the deliverable D.4.2. The key software integration innovations underlying all UC0, UC1, UC2 are the highly scalable TokenQueue load-balancer described in the deliverable D4.1 and its DockerSpaces implementation for robust scaling in CPU and GPU clouds described in the deliverable D4.3. SELMA platform has been already noticed and featured on EU Innovation Radar website[1].

Since the overall SELMA platform architecture has been stable since the interim report D4.3, to avoid duplication, in this deliverable we focus on evolution of the SELMA Use Case 0 from a project-internal NLP software testing environment into the SELMA Open-Source Software (SELMA OSS) release[2], which is published also in EU Common Language Resources and Technology Infrastructure (CLARIN) https://www.clarin.eu/ under handle http://hdl.handle.net/20.500.12574/97 and thus made available far beyond the SELMA project itself.

Various color-coded versions of SELMA UC0 illustrate the changing landscape of the open-source NLP software becoming available throughout the course of the SELMA project, as it affected what NLP functions could be solved with external off-the-shelf components, and where SELMA-developed NLP modules provide superior performance. In this way SELMA UC0 served also as a gateway between the NLP software features developed inside and outside the SELMA project.

[1] https://innovation-radar.ec.europa.eu/innovation/48642

[2] https://github.com/SELMA-project/UC0-OpenSource

The central trend observed during the SELMA project was continuous release of high-quality open-source NLP software by large US companies like OpenAI (Whisper multilingual ASR with full punctuation) and Meta (multilingual wav2vec ASR, direct machine translation between 100 languages, TTS for more than 100 languages, Large Language Model “Llama”) as part of their mutual competition for dominance in the arena of the powerful AI. The integration flexibility of SELMA UC0 allows leveraging not only the research done within SELMA but also these open-source software releases.

2. “White” Version of Use Case 0: SELMA NLP Service Orchestration

The initial “White” version of Use Case 0 (shown in Figure 1) was described already in the deliverable D4.2, here we provide only a brief recap.

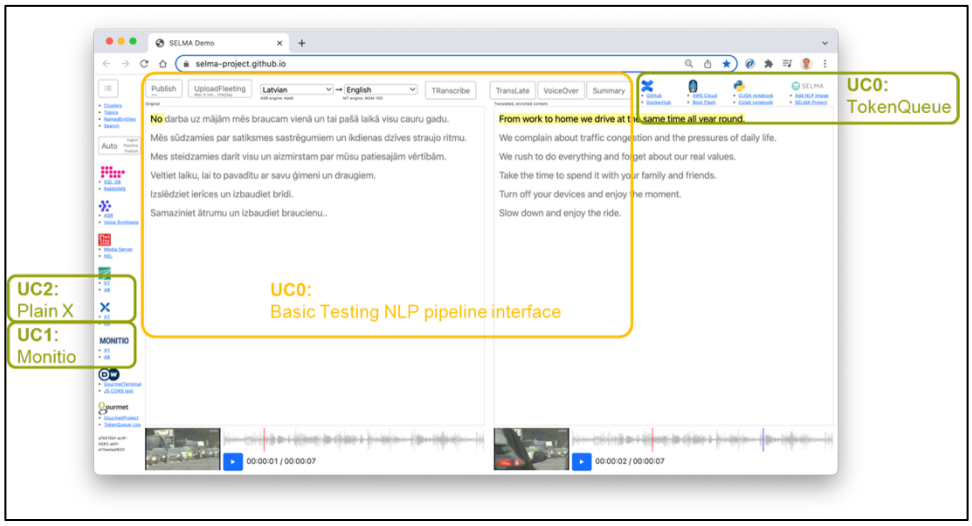

Figure 1 *”White” version of the SELMA Basic Testing and Configuration Interface (UC0)*

The “white” version at the beginning of the SELMA project provided the Basic Testing and Configuration Interface for testing, deployment, scaling and monitoring of the NLP services developed within the SELMA project WP2 and WP3. The NLP worker deployment initially used the TokenQueue mechanism (described in D4.1) to deliver highly scalable SELMA NLP Service for the primary Use Cases UC1, UC2.

The “white” UC0 is integrated with NLP-pipeline execution orchestrator Maestro (also described in D4.1) shared with the UC1 and UC2. Maestro Orchestrator served as a gateway between all three Use Cases UC0, UC1, UC2 allowing them to share the same NLP worker pool defined in the TokenQueue.

Technically the “white” UC0 software release consists of three components:

- TokenQueue backend handling the scalable deployment of the SELMA NLP workers through the long-polling TokenQueue protocol. TokenQueue is controlled by the dynamically updated list of available NLP workers.
- The actual NLP workers deployed on various servers within the SELMA Consortium and accessed in the uniform manner via the TokenQueue backend. The list of deployed workers includes both SELMA WP2, WP3 outcomes as well as other open-source NLP workers such as neural machine translation from GoURMET H2020 project Deutsche Welle was participating.
- Highly scalable UC0 GUI frontend (Figure 1) served directly from the GitHub repository of the SELMA project (https://selma-project.github.io/) implementing the basic SELMA NLP pipeline for testing, monitoring and configuration purposes.

In the backend "white" UC0 version was running a wide range of NLP services, including:

- About 30 bilingual translation services from the GoURMET H2020 project Deutsche Welle was participating,
- Multiple ASR versions based on Kaldi and wav2vec technologies available from the SELMA project partners at that time.
- The first Brazilian Portuguese TTS system trained in the SELMA project by the Avignon University using Deutsche Welle Brazilian broadcast recordings. Later in the project this was extended with Urdu TTS system trained in the SELMA project by the Avignon University using Deutsche Welle Urdu broadcast recordings. Finally, Latvian TTS system was trained in the similar manner in the SELMA project by IMCS using Latvian blind people library audiobooks.
- Named Entity Recognition and Linking along with semi-manual continuous massive stream learning of varied spellings of these entities was included using the PiniTree software from IMCS. Despite being integrated into the initial UC0 "white" version, this feature faced two obstacles – it was considered violating GDPR and ethical guidelines for a cloud-public open-source software; and the accurate semi-manual operation mode of PiniTree was considered inferior to alternative less-accurate fully-automatic LLM based approaches developed within the project. As the result, this feature has been removed from the UC0 "white" version and replaced by the red/yellow/green semaphore icon pointing to the follow-up versions of UC0 described below.

- NER modules trained in SELMA by Priberam on DW and LETA data for Russian, Ukrainian, Latvian, Dutch, Turkish.

3. “Yellow” Version of Use Case 0: FrontEnd NLP Pipeline and DockerSpaces Backend

The second “Yellow” version of the UC0 (shown in Figure 2) corresponds to the robust DockerSpaces CPU and GPU cloud implementation of the initial TokenQueue load-sharing mechanism as described in the deliverable D4.3.

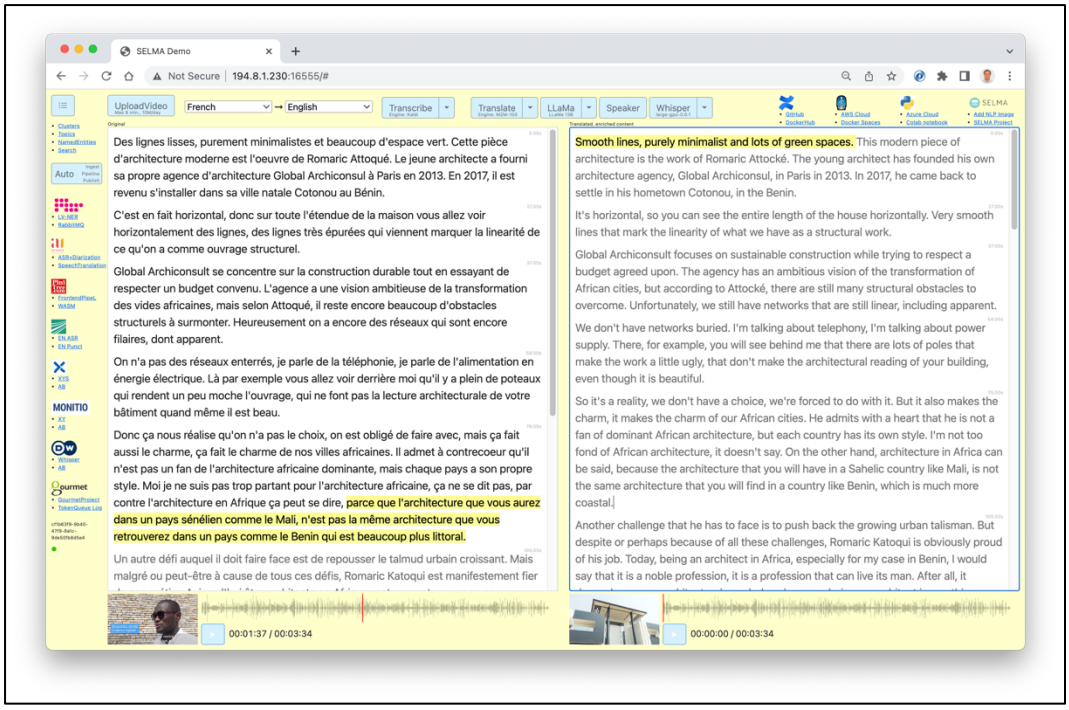

Figure 2 *”Yellow” version of the UC0: frontend NLP pipeline with universal DockerSpaces backend*

Although visually the “yellow” version is nearly identical to the “white” version described above, the key NLP software integration innovation there is that the entire NLP processing pipeline now is moved from the backend to the frontend GUI JavaScript running directly in the end-user web browser.

This had far-reaching consequences in the ease of adding new NLP tools to the UC0 GUI. If before (in the “white” version) integration of each new NLP tool required both frontend and backend code changes and support for new REST API call between them, then in the “yellow” version universal DockerSpaces backend allowed calling auto-scaled NLP workers directly

from the GUI frontend JavaScript code. Consequently, the “yellow” version of UC0 saw a record number of new and experimental NLP functions integrated.

The key innovative NLP techniques initially integrated and tested in the “yellow” UC0 version (and later integrated into the main use cases UC1 and UC2, as well as in the Podcast Creation Use Case) were:

- Large Language Model (LLM) “LlaMa” and its instruction tuned version “Alpaca”. At the time OpenAI ChatGPT has recently been released and it was essential to get similar functionality inside the SELMA NLP toolbox for news content abstractive summarization without sending the private user data to the OpenAI API running in USA. Eventually this functionality got integrated in the SELMA Podcast Creation Use Case. The key limitation of LlaMa and other open-source LLMs at the time was their limited language coverage. Only towards the very end of the SELMA project (March 17, 2024) xAI released the first open-weights model Grok with true multilingually for 200 languages, successfully tested by IMCS.
- Text To Speech (TTS) for English with arbitrary voice provided through the brief 5-minute audio file of the desired voice (rather than 30 hours required before to train the Brazilian Portuguese TTS integrated in the “white” version of UC0).
- Voice conversion of any audio file in any language with arbitrary voice provided through the brief 5-minute audio file of the desired voice. This feature eventually triggered an ethics alarm (possibility to create a fake audio in the voice of the person, who never spoke this text) and was removed in the later versions of the UC0.
- Multilingual OpenAI Whisper ASR and punctuation model with integrated language detection for comparison with the Kaldi and Wav2vec ASR models and separate punctuation post-processing available in the “white” version of UC0. Whisper quality and language coverage was stunning and shifted SELMA project ASR research from finetuning only wav2vec ASR models to finetuning also Whisper ASR models.
- Direct translation engine between 100 languages M2M-100 from Meta was added as alternative to GoURMET H2020 project bilingual translation engines.
- Speaker diarization system developed in SELMA by Fraunhofer.
- Speech-to-speech translation between French and English developed in SELMA by the University of Avignon.

The DockerSpaces universal backend behind the "yellow" UC0 version has been open sourced on its own at the address https://github.com/SELMA-project/docker-spaces. The **scalability tests to reach 10M documents served per day** have successfully been conducted using the M2M-100 language translation NLP module as the reference application. The results showed that on the Apple ARM M2pro CPU with 1Gbps connection 10M HTTP load-sharing redirect requests can be served in 3 hours. Meanwhile M2M-100 translation module on Nvidia DGX server with 8xA100 GPU cards and 128 CPU cores for translation achieved 240 docs/min using 8xGPU and 30 docs/min using only 128xCPU cores; on Apple ARM M2pro CPU with 10 cores we achieved 3 docs/min for translation, which is similar to Intel/AMD x86-64 CPU cores in DGX, but with 6x less power consumption. The power consumption matters, if the system indeed would have to be run at 10M docs/day translated, as it would require 29 Nvidia DGX servers with 8xA100 GPU cards and 128 CPU cores running in parallel, consuming 5KW x 29 x 24h = 3,480MWh/day. At 0.30 EUR per 1 KWh it would cost 1044 EUR/day in electricity alone. The ability to reduce power consumption with Apple ARM M2 chips 6x would reduce the running cost to 174 EUR/day.

The actual document translated in the tests was a two-paragraph Franz Kafka text, which got translated from English to German:

One morning, when Gregor Samsa woke from troubled dreams, he found himself transformed in his bed into a horrible vermin. He lay on his armour-like back, and if he lifted his head a little he could see his brown belly, slightly domed and divided by arches into stiff sections. The bedding was hardly able to cover it and seemed ready to slide off any moment. His many legs, pitifully thin compared with the size of the rest of him, waved about helplessly as he looked.

"What's happened to me?" he thought. It wasn't a dream. His room, a proper human room although a little too small, lay peacefully between its four familiar walls. A collection of textile samples lay spread out on the table—Samsa was a travelling salesman—and above it there hung a picture that he had recently cut out of an illustrated magazine and housed in a nice, gilded frame. It showed a lady fitted out with a fur hat and fur boa who sat upright, raising a heavy fur muff that covered the whole of her lower arm towards the viewer.

4. “Green” Version of Use Case 0: SELMA Open-Source Candidate

The third “Green” version of the UC0 (shown in Figure 3) was intended as the SELMA Open-Source Software (SELMA OSS) release and technically is the same “yellow” version of the UC0 with frontend NLP pipeline and universal DockerSpaces cloud-scalable backend, only with unnecessary functionality and clutter removed from the GUI to make it more appealing as a cloud-hosted application for multilingual video subtitling, translation and voice-over also outside the SELMA project.

For the dissemination purposes of this “green” intended SELMA OSS release, a video tutorial was created by Deutsche Welle, available on YouTube at https://youtu.be/r4nsumsnR5M.

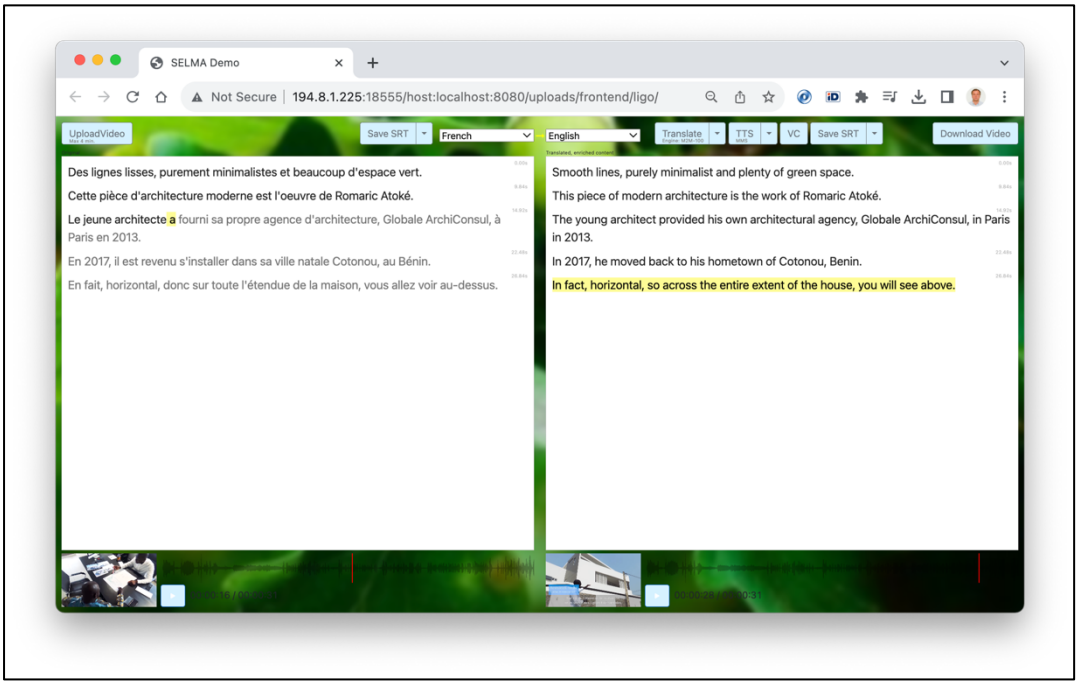

Figure 3 *”Green” version of the UC0: SELMA OSS candidate with scalable DockerSpaces backend*

Unfortunately, this candidate version of SELMA OSS software hit a roadblock from the SELMA project ethics review – it had three ethics flaws:

1. Since the “green” SELMA OSS allows any user on the Internet to upload and process any video, audio or text content, this would lead to that content being stored in the SELMA consortium servers, along with the IP address of the user visible in the server logs. This causes two-fold concern of users inadvertently uploading personally identifiable private content about themselves or other persons, or otherwise illegal content.
2. The Voice Conversion functionality (button “VC” in Figure 3) was completely unacceptable. It allowed voice conversion for any audio file in any language with arbitrary voice provided through the brief 5-minute audio file of the desired voice. This feature triggered an ethics alarm because of the possibility to create a fake audio in the voice of the person who never spoke this text. This feature was described as “stealing the voice” of another person as it could be used even for criminal activities like extortion through fake financial help phone-calls from the voices of close relatives.
3. Finally, the Latvian TTS voice was trained from the audio-books available in the Latvian blind people audio-library with the permission of that library, so that it could be used to produce more audiobooks for the blind in Latvia. The SELMA project trained synthetic voice of the popular Latvian actor, who used to read these audiobooks, turned out to be very high quality, hardly distinguishable from the actual Latvian actor, who described the situation as “his soul has been stolen” - the very specific and likable way of speaking and reading, what makes him so popular and also so trusted to read only high quality content. He was very concerned that some low quality or harmful content could be now synthesized in his voice.

Due to these triple ethical concerns, the final SELMA OSS software had to be redesigned almost from scratch – the mere dropping of all three features would make the whole UC0 useless.

5. "Red" Version of Use Case 0: Final SELMA Open-Source Software

The "Red" version of the UC0 (shown in Figure 4) is the final SELMA Open-Source Software (SELMA OSS) release[3]. Visually it looks like the "green" version with only the controversial "VC" Voice Conversion button removed. But under the hood, everything has changed to make this a completely private version of UC0 running entirely on the end-user computer (both frontend and backend), so that no data ever leaves the end-user computer.

The most crucial technical innovation in the "red" UC0 version is related to getting rid of the CUDA GPU accelerator requirement for the most NLP modules included in the UC0 to run at usable speed, as we cannot expect every end-user to possess a power-hungry, noisy and expensive CUDA GPU in his laptop of office computer. This would have been a dead-end a few years ago, but we found a solution with the latest breed of CPUs (like Apple ARM M1, M2, M3) incorporating neural acceleration hardware similar to GPUs to accelerate inference for neural network-based NLP components.

[3] https://github.com/SELMA-project/UC0-OpenSource

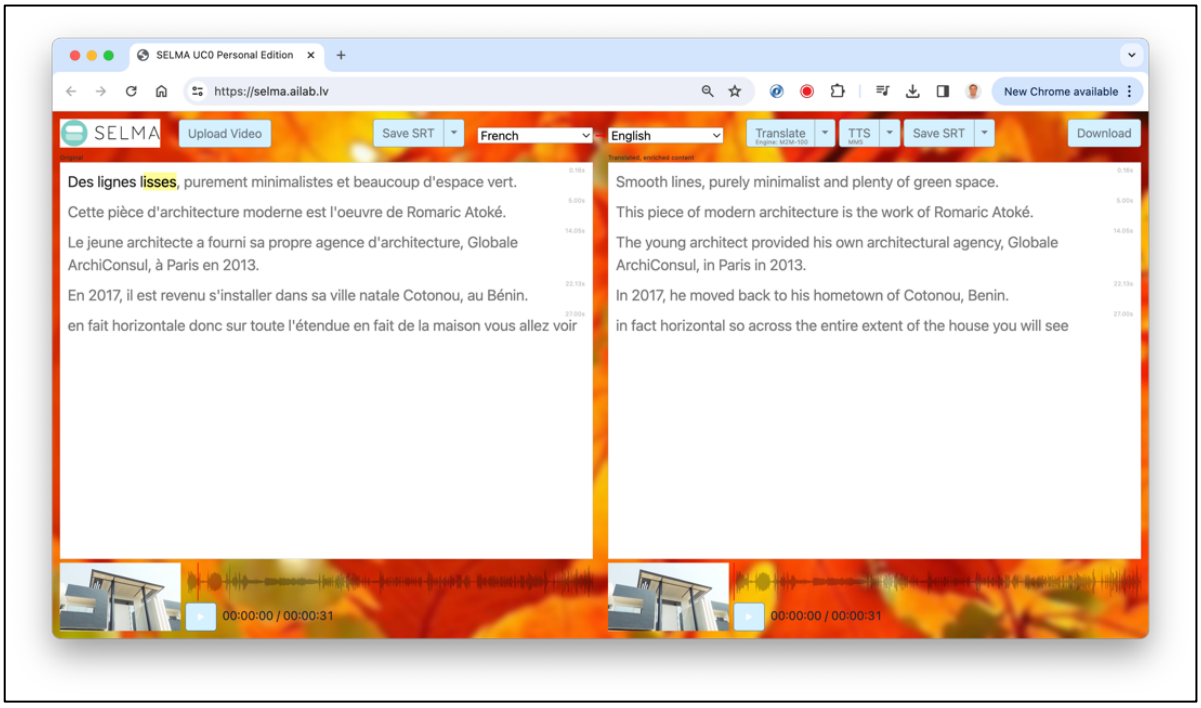

Figure 4 *"Red" version of the UC0: final SELMA OSS with statically linked backend for Apple MacOS (ARM M1, M2, M3 or Intel) or Linux (x86-64) end-user computer*

The latest technological innovation in WP4 - the private "red" version of UC0 branded also as "SELMA OSS" version solves privacy issues of the "green" version and enables unlimited scalability by running entirely locally on the end-user computer. Scalability is further improved by the advent of the 6x more energy efficient Apple ARM-CPUs M1, M2, M3 in personal computers – although SELMA OSS runs also on x86-64 architecture in Linux, only it is slower and less power-efficient there.

The SELMA Open-Source Software (OSS) offers effective means to test and compare the performance of various language models used in multilingual media monitoring and content production. The SELMA OSS Platform (also referred to as Use Case 0, UC0, or The Basic Testing and Configuration Interface) provides:

- automatic speech recognition (ASR) from audio/video files,
- punctuation and capitalization of the transcribed text,
- machine translation (MT) into a target language,
- text-to-speech synthesis (TTS) and synchronous voice-over generation.

To provide this functionality, the demonstrator release uses these multilingual open-source models: OpenAI Whisper (ASR), Meta MMS (TTS, ASR), Meta M2M-100 (MT). Thus, it facilitates easy access to such open large language models.

The SELMA Platform can be used not only by developers to combine and test alternative language models before they are integrated into the end-user applications – it can also be used as an entry-level application by journalists and media producers themselves to transcribe their recordings, generate subtitles and voice-over, or to generate a podcast from an input text.

The demonstrator of the SELMA OSS Platform at address https://selma.ailab.lv/ does not require registration and authentication nor does it store any content, original or generated, after the session is closed by the user; demonstrator will be maintained there for 2 years after the end of the SELMA project. Meanwhile the SELMA OSS release itself is published for download both on GitHub https://github.com/SELMA-project/UC0-OpenSource and in EU Common Language Resources and Technology Infrastructure (CLARIN) under the persistent handle http://hdl.handle.net/20.500.12574/97 indefinitely.

6. Conclusion

Besides presenting the final SELMA platform release with full continuous massive stream learning capabilities, this deliverable focuses on testing Use Case 0 introduced in the SELMA project and its various versions on the way towards the final "red" SELMA Open-Source Software release. The key takeaway from the two branches of SELMA software integration efforts is that GDPR and ethical guidelines are stricter for open-source and public-cloud deployed services like UC0, than they are for closed-source commercial software pursued in UC1 and UC2. With a technologically innovative approach, the SELMA project was nevertheless able to release a personal edition of SELMA OSS., open to the public at large and complying to the current GDPR and ethical guidelines.