Research and Innovation Action (RIA) H2020 – 957017

Stream Learning for Multilingual Knowledge Transfer

https://selma-project.eu/

D5.3 Final Evaluation Report

Work Package	5
Responsible Partner	Deutsche Welle
Author(s)	Peggy van der Kreeft
Contributors	Andreas Giefer, Hala Attig, Kay Macquarrie, all partners
Reviewer	Guntis Barzdins
Version	V1.0
Contractual Date	31 March 2024
Delivery Date	28 March 2024
Dissemination Level	Public

Version History

Version	Date	Description
0.1	22/02/2024	Initial Table of Contents (ToC)
0.2	05/03/2024	Initial Input
0.3	12/03/2024	Input from Technical Partners
0.4	20/03/2042	Internal Review
1.0	26/03/2024	Finalization and Submission

Executive Summary

Evaluation was central to SELMA's activities. It has taken the ambitious technological research and prototype development to a next level by determining its added value, its strengths and weaknesses and room for improvement, and - last but not least - its usefulness for the media world, our focus user group, for the three use cases, multilingual media monitoring and news production, as well as the open-source SELMA system.

This document provides a final update of the evaluation efforts within the SELMA project, according to the previously established Evaluation Plan. It describes what has been done in terms of assessing the targeted platforms and use cases, as well as individual components. It showcases what technology partners and user partners have done and have collaborated on in this respect.

This report describes the overall progress on evaluation in section 2. Section 3 refers to the work done on technical testing by the technology partners, on the 13 SELMA components and the three integrated platforms (SELMA Open-Source Platform, Monitio and plain X).

Table of Contents

Table of Figures

Table of Tables

1. Introduction

The SELMA project has developed a StrEam Learning for Multilingual knowledge-trAnsfer (SELMA) platform, integrating different NLP (natural language processing) tools. More details on the objectives were given in D5.1 - Evaluation Plan.

This document focuses on the execution of the evaluation plan for the SELMA project, as described in D5.1 - Evaluation Plan. We will show progress in the evaluation in Y3 done and a general overview of what has been achieved on evaluation at different levels:

- individual components
- integrated platform and demonstrators
- targeted use cases and use case applications

2. Evaluation Plan and Progress Made

This section provides a broad overview of the evaluation activities in Year 3, as well as a final overview and status.

The principal objective is to ultimately develop technologies and tools that are stable, easy to use, flexible and expandable.

As outlined in the Evaluation Plan, SELMA partly builds upon proven prototypes, in particular the SUMMA platform for monitoring and the plain X platform for content creation and adaptation. It extends these with continuous transfer learning capability from external data streams and user feedback, resulting in a system that becomes better with increased use, capable of ingesting massive amounts of different sources (news, internet feed, social media, etc.), and produce well-organized and topic-driven information that facilitates the propagation of key information to the end users.

The main evaluation objectives we are pursuing are listed in the following table, with an indication of those objectives that have been evaluated during the reporting period.

#	Evaluation Objective	Achieved
1	Evaluate the outcomes of the novel methods for training (and updating) machine learning/deep learning models for multiple speech and language tasks continuously.	Y
2	Evaluate and benchmark the outcomes of the newly developed unsupervised multilingual language models for all 30+ project languages.	Y
3	Evaluate the improvement of downstream tasks like entity recognition and linking, topic labelling, clustering, transcription, abstractive news summarization, automatic post-editing in all 30+ languages.	Y
4	Evaluate different clustering algorithms.	Y
5	Evaluate outcomes of knowledge transfer across tasks in situations with asymmetrical amounts of resources between languages and tasks, particularly low resource languages.	Y
6	Evaluate the newly developed data analytics methods and visualizations for improving the readability and access to information in order to boost and	Y

	facilitate the decision-making process of media monitoring analysts and any global end-user in terms of accuracy and usefulness.	
7	Evaluate functionality, usability and user acceptance for media monitoring workflow.	Y
8	Evaluate functionality, usability and user acceptance of the multilingual content production workflow, particularly the multilingual transcription and translation models trained within the SELMA platform to enable an editorial production and content re-use workflow for 30 languages.	Y
9	Evaluate overall media monitoring workflow for analytics for decision-making by media professionals	Y
10	Monitor, validate and evaluate the outcome of the newly developed user feedback input and self-learning workflow for the improvement of the deep-learning model.	Y
11	Evaluate whether the usage of the integrated workflows enabled by the SELMA platform will measurably improve the ease of multilingual content monitoring and creation. Evaluate the overall acceptance of the novel tools and workflows.	Y

Table 1 *Evaluation Objectives*

Basic Component Overview

The following table provides an updated list of components developed and assessed within the project.

Component	Partners involved in development and assessment
1. Automated Speech Recognition (ASR)	LIA, FhG, IMCS, Priberam, DW/users
2. Machine Translation (MT)	LIA, FhG, IMCS, Priberam, DW/users
3. Summarization	Priberam, IMCS, DW/users
4. Named Entity Recognition (NER), Named Entity Linking (NEL)	LIA, Priberam, IMCS, DW/users

5. Post-editing	LIA, FhG, Priberam, IMCS, DW/users
6. Clustering	Priberam, FhG, IMCS, DW/users
7. Topic detection	Priberam, FhG, IMCS, DW/users
8. Speech synthesis	LIA, FhG, Priberam, IMCS, DW/users
9. Story segmentation	FhG, Priberam, IMCS, DW/users
10. Punctuation & TrueCasing	FhG, Priberam, DW/users
11. Speaker Diarization	FhG, Priberam, DW/users
12. Speaker Recognition	FhG
13. Graph Orchestrator platform	Priberam
14. Monitio platform*	Priberam, DW/users
15. plain X platform*	Priberam, DW/users
16. SELMA OSS platform*	IMCS, DW/users
*full platforms with integrated components	

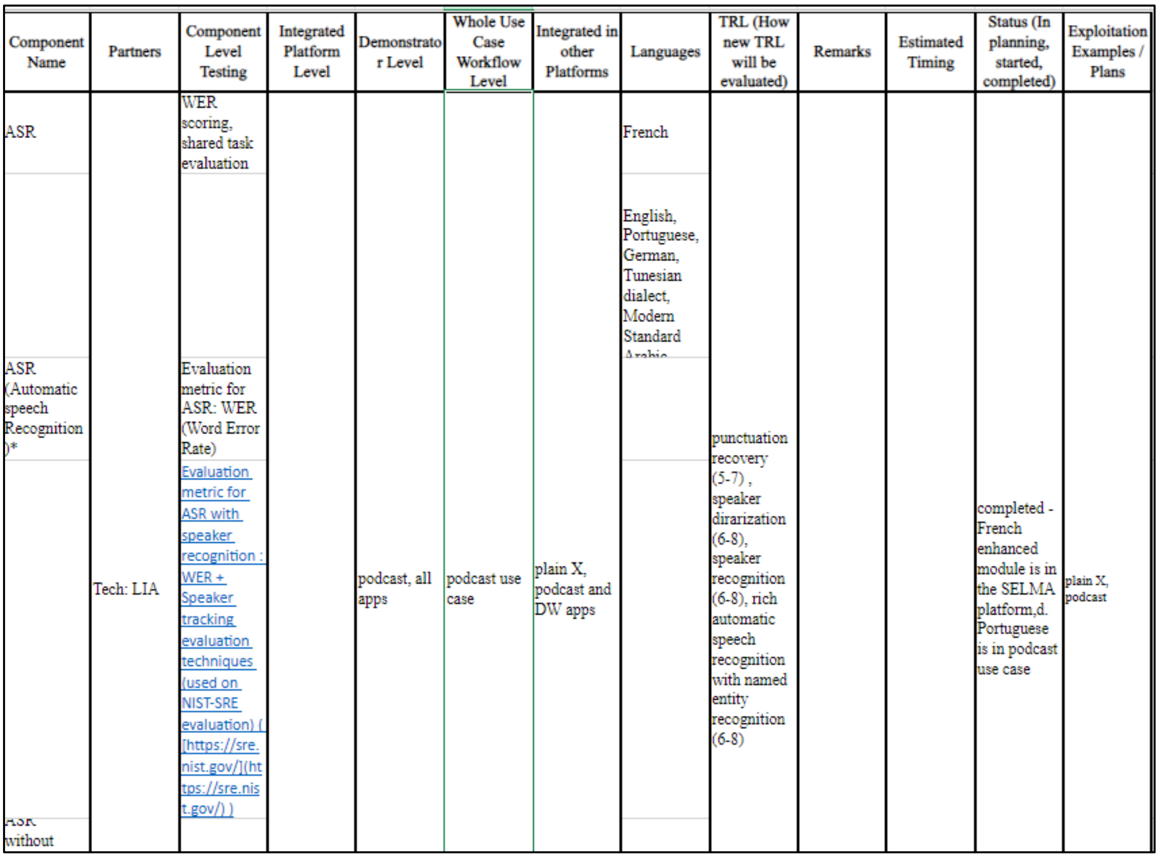
Table 2 *Basic Component Overview*

The expected TRL (Technology Readiness Level) for each of the technologies or components has been achieved or exceeded, as can be seen in table13 in D6.6 (Final Impact Report), section 5.6.

Detailed Component Evaluation Tracker

The Evaluation Excel Sheet below shows an updated status and keeps track of the evaluation activities at the different levels and by the different consortium partners. This table served as our main evaluation tracking tool. It lists the components developed and evaluated within the project, with details on what aspects are the focus per partner and what kind of evaluation is planned. It functioned as a live document and was continuously updated and expanded throughout the project. It contributed to the Basic Component Overview as displayed above.

Below is a screenshot of the updated Excel sheet containing the Evaluation Tracking sheet:



Component Name	Partners	Component Level Testing	Integrated Platform Level	Demonstrator Level	Whole Use Case Workflow Level	Integrated in other Platforms	Languages	TRL (How new TRL will be evaluated)	Remarks	Estimated Timing	Status (In planning, started, completed)	Exploitation Examples / Plans
ASR	Tech: LIA	WER scoring, shared task evaluation		podcast, all apps	podcast use case	plain X, podcast and DW apps	French	punctuation recovery (5-7) , speaker dirarization (6-8), speaker recognition (6-8), rich automatic speech recognition with named entity recognition (6-8)			completed - French enhanced module is in the SELMA platform,d. Portuguese is in podcast use case	plain X, podcast
							English, Portuguese, German, Tunesian dialect, Modern Standard Arabic					
ASR (Automatic speech Recognition)*		Evaluation metric for ASR: WER (Word Error Rate)										
		Evaluation metric for ASR with speaker recognition : WER + Speaker tracking evaluation techniques (used on NIST-SRE evaluation) (https://sre.nist.gov/)										
ASR without												

Figure 1 *Screenshot of Updated Excel Sheet with Detailed Evaluation Plan*

Test Users and User Group

The test user group contains primarily DW media professionals, selected based on their language proficiency and use case interest. Thus, specific groups have been formed for NLP benchmarking, media monitoring, news production, diversity analysis, podcasting and speech synthesis, for instance. We involved users from the Arabic, Serbian, Turkish, Brazilian, Indonesian, Kiswahili departments, as well as the Archive and Documentation Center, the Business Department and the Technology Strategy Department. In addition, a test team consisting of Priberam clients, such as EMBRAER, AICEP and LUSA, have also evaluated the SELMA enhancements to the integrated platforms Monitio and plain X.

The project has also set up a User Group of 31 members, from the EBU, SWR, ARTE, RAI, BBC, EuroNews, Prisa, EMBRAER, AICEP, and research organizations such as the University of Edinburgh and the University of Tilburg. The members of the User Group and Advisory Board, a subset of the User Group, were informed about how the project advanced, had access to some of the platforms, participated in the meetings and some actively participated in the trials.

Three User Events were organized in the course of the project during which user feedback was gathered. The first (hybrid) User Day at Deutsche Welle took place in Bonn on 12 October 2022. This hybrid event included presentations of the SELMA progress over the first project half. We had up to 60 participants in the remote sessions (presentations and panel discussion) and 20 in the onsite workshop and demo sessions in the afternoon. The remote sessions allowed members of the audience, including those from the official SELMA User Group, to comment, ask questions, and provide suggestions. This opened up the view and allowed for some new directions. More details on the first SELMA User Day can be found in D6.4 - Interim Impact Report.

A second SELMA User Day was organized in the framework of the Festival IA Conference, coordinated by LIA. It was a two-day on-site event in Avignon on 14-15 November 2023 and included presentations as well as demos for participants from the research community, government and industry. Details can be found in D6.6 - Final Impact Report.

Towards the end of the project, a final User Event was organized on 21 March 2024, this time fully virtual. The event with almost 40 participants including members of the user group allowed us to obtain feedback from our user group as to the prototypes and the applications built upon the SELMA technologies. This is further described in D6.6 - Final Impact Report.

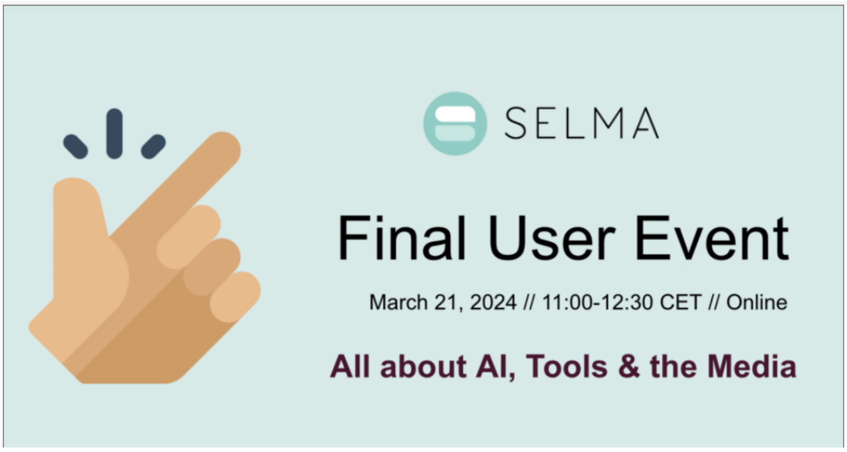

SELMA
Final User Event
March 21, 2024 // 11:00-12:30 CET // Online
All about AI, Tools & the Media

***Figure 2** Invitation Banner to the Final User Event*

3. Technical Testing

This section addresses technical testing for individual components as well as integrated platforms and demonstrators. We mainly list the components and platforms that are subject to technical testing.

Details on specific technical testing are reported in the respective technical deliverables:

- D1.4 Final Prototype Report
- D2.7 Final progress report on continuous massive stream learning
- D2.8 Final release of continuous massive stream learning tools
- D3.7 Final report on speech and natural language processing
- D3.8 Final release of speech and natural language processing tools
- D4.4 Final platform release with full continuous massive stream learning capabilities

SELMA NLP components developed by the University of Avignon (LIA), Fraunhofer (FhG), Priberam, and IMCS are primarily tested on their own, sometimes with a special UI. The purpose is to validate that the software of the component performs as expected. This first-level testing is done by the developing partner and precedes integration testing.

Evaluation of the 16 components listed in Table 2 - Basic Component Overview started in Year 1, intensified in the second project year, and was finalized in year 3. An overview is provided below, with full details available in the respective technical deliverables.

- Component 1: ASR – D3.7 Final report on speech and natural language processing; D3.8 Final release of speech and natural language processing tools; D4.4 Final platform release with full continuous massive stream learning capabilities
- Component 2: MT – D3.7 Final report on speech and natural language processing; D3.8 Final release of speech and natural language processing tools; D4.4 Final platform release with full continuous massive stream learning capabilities
- Component 3: Summarization – D2.7 Final progress report on continuous massive stream learning; D2.8 Final release of continuous massive stream learning tools; D4.4 Final platform release with full continuous massive stream learning capabilities

- Component 4: NER/NEL – see D2.7 Final progress report on continuous massive stream learning; D2.8 Final release of continuous massive stream learning tools; D3.7 Final report on speech and natural language processing; D3.8 Final release of speech and natural language processing tools
- Component 5: Post-editing of ASR and MT – see D3.7 Final report on speech and natural language processing; D3.8 Final release of speech and natural language processing tools
- Component 6: Clustering – see D2.7 Final progress report on continuous massive stream learning; D2.8 Final release of continuous massive stream learning tools
- Component 7: Topic Detection – see D2.7 Final progress report on continuous massive stream learning; D2.8 Final release of continuous massive stream learning tools
- Component 8: Speech Synthesis – D3.7 Final report on speech and natural language processing; D3.8 Final release of speech and natural language processing tools; D4.4 Final platform release with full continuous massive stream learning capabilities
- Component 9: Story Segmentation – D2.7 Final progress report on continuous massive stream learning; D2.8 Final release of continuous massive stream learning tools; D4.4 Final platform release with full continuous massive stream learning capabilities
- Component 10: Punctuation & TrueCasing – D3.7 Final report on speech and natural language processing; D3.8 Final release of speech and natural language processing tools
- Component 11: Speaker Diarization – D3.7 Final report on speech and natural language processing; D3.8 Final release of speech and natural language processing tools
- Component 12: Speaker Recognition – D3.7 Final report on speech and natural language processing; D3.8 Final release of speech and natural language processing tools
- Component 13: Graph Orchestrator platform – D2.7 Final progress report on continuous massive stream learning; D2.8 Final release of continuous massive stream learning tools; D4.4 Final platform release with full continuous massive stream learning capabilities
- Component 14: Integrated Monitio platform – see D1.4 Final prototype report; D4.4 Final platform release with full continuous massive stream learning capabilities
- Component 15: Integrated plain X platform – D1.4 Final prototype report; D4.4 Final platform release with full continuous massive stream learning capabilities

- Component 16: Integrated SELMA OSS – D1.4 Final prototype report; D4.4 Final platform release with full continuous massive stream learning capabilities

The final status per component can be found in more detail in the deliverables listed above.

4. User Evaluation

User Evaluation has taken place at several levels. We evaluate the individual components in close cooperation with the developers, at platform level (SELMA OSS, Monitio and plain X) with new SELMA components integrated through their new or enhanced functionalities, as well as through the use cases and use case applications.

In Y3, Usability testing - pursuing the five E's: effective, efficient, engaging, error-tolerant and easy to learn (https://www.wqusability.com/) - was the focus, both at platform and use case level.

Thus, the SELMA OSS, Monitio and plain X platforms with new or updated SELMA features and functionalities, and the use case applications, including podcasting and diversity applications, were at the core of the user evaluation efforts. Concrete and measurable feedback on usability was obtained through testing, regular use, user observation, questionnaires, and interviews.

A major part of the assessment is through interaction with the users actually working with the systems and providing feedback, then changing settings or suggesting enhancements, and implementing them in the platforms and applications. Some of these interactions or suggestions are described in the sections below.

4.1 Integrated SELMA OSS

In **Year 1**, it was decided to add an extra use case, i.e., UC0, enabling SELMA's open-source software (OSS) components to be made accessible via a unified API under the umbrella of use case 0. Several UIs and tools were made available through the OSS platform and initial testing of those UIs and tools was done in the first year.

This API offers the possibility to evaluate individual components through command-line tools and more advanced user applications.

In **Year 2,** improvements on the OSS tools were further evaluated, including the basic transcription-translation-speech synthesis workflow in this UI. New NLP engines were added, including the speech translation developed by LIA and the DW customized Brazilian voices. This workflow was tested from a user point of view, in terms of ease of use, easy access, speed and language coverage. It is considered useful as a low-threshold tool for demoing this kind of workflow and a basic view of automation of NLP processes, but is limited for choice of engines (only OSS engines). That is acceptable because of the purpose of the OSS, while a more sophisticated platform with a variety of additional commercial engines is available through plain X.

In addition, two components of the SELMA OSS were evaluated for user applications using the command-line API.

First, low-resourced language translation modules from the GoURMET – Global Under-Resourced Media Translation – a Horizon 2020 project – in which Deutsche Welle participated and which ended in March 2022 - were made available on-demand through SELMA's DockerSpaces orchestrator. This enabled DW as end user to try out and put the MT engines coming from this EU project (Grant agreement 825299) to use. All 16 engines were integrated into the SELMA OSS. The engines were installed in SELMA as dockerized modules.

A command-line tool allowed us to put the system through its paces by sending repeated translation requests for a variety of source and target languages. The following figure shows the number of words per seconds for various source/target language pairs that the system managed to translate. Running the script repeatedly allowed us to test the orchestrator's stability and dependability, while highlighting areas that required improvements.

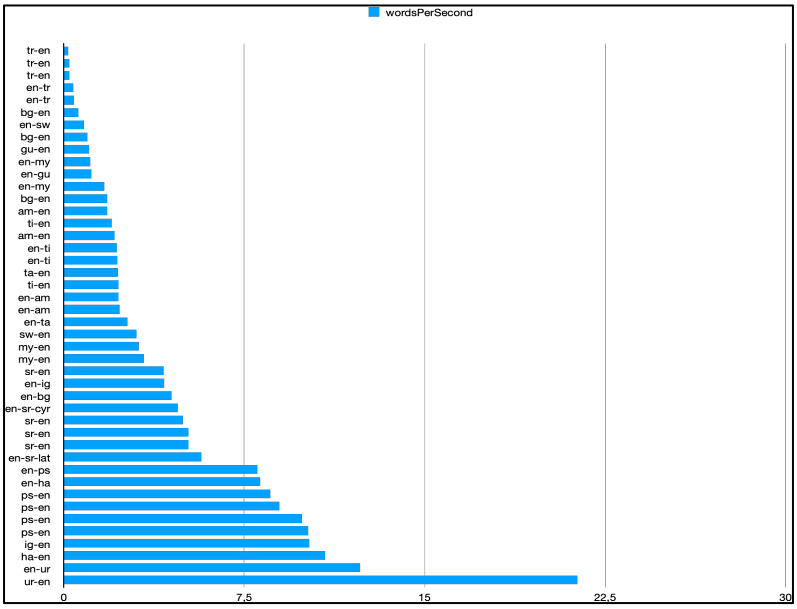


__Figure 3__ Benchmarking of GoURMET modules with SELMA OSS

The GoURMET evaluation was extended with a Word Cloud analysis application. This iOS application was developed by the DW SELMA project team. It starts with a batch translation of a selected set of texts, translated by selected engines, in this case the GoURMET dockerized modules in SELMA OSS. The translations were automatically retrieved and a subsequent analysis for content focus was done, resulting in a Word Cloud output. This allows us to do automated batch processing of selected source documents for a topical content analysis.

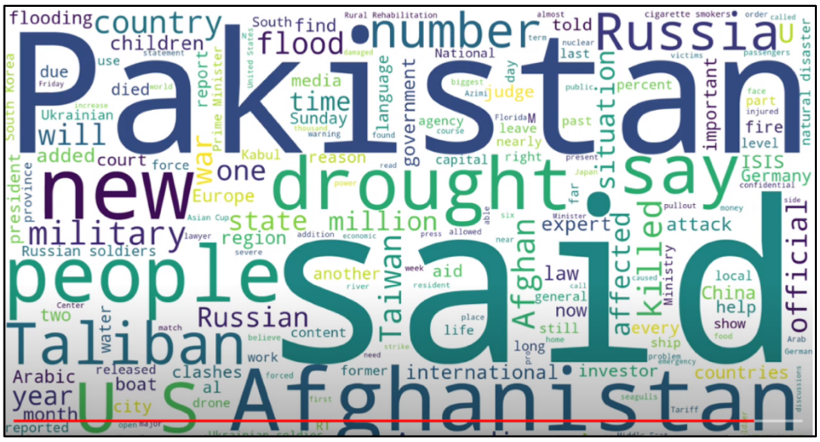


__Figure 4__ Word Cloud application in SELMA OSS

The following screencast shows how the Word Cloud application works:

https://www.youtube.com/watch?v=ejNTUd9Tda0

Second, the Podcast creator application uses a similar DockerSpaces API to convert news bulletin texts to Brazilian speech. Again, this allowed us to highlight areas where either the text-to-speech Docker module needed updating or the Orchestrator's stability required improvements. More details on the Podcast creator evaluation are given in section 4.5.

In **Year 3**, we continued to try out the orchestration in UC0 and the SELMA Dockerspaces, did a comparative analysis and looked for additional applications.

We looked at the SELMA OSS from the user point of view, for potential use. We accessed it in two ways: through the UI (https://selma.ailab.lv/#) and through the API.

In terms of usability, for the regular, non-technical user, the overall OSS UI is quite technical and can be somewhat overwhelming, as it is a gateway to a large variety of services. It makes a lot of functionality accessible to the user. It is good to have such an overview, but it is not the everyday UI a professional editor is used to. However, that is not the primary objective, it is not set up to be a permanent, widely used UI for a professional environment. It is an example of what the system could look like if the open-source software is installed by interested users.

The core NLP process providing a very basic transcription-translation-voice-over workflow serves its purpose: it is a simple, extremely transparent, UI, running these three processes, without need for further explanation or training.

We evaluated the overall NLP processes through this UI (selma.ailab.lv), and compared it with the plain X workflows.

Interface

In the SELMA OSS, all functionalities are available directly in the main interface. Users upload a video file using the Upload Video button. To translate a text, they need to select the right source language and target language. To select a specific engine for translation, the user needs to click on the arrow next to the Transcription button to get a list of available engines. The platform also uses knowledge-specific keywords (for example, TTS for text-to-speech).

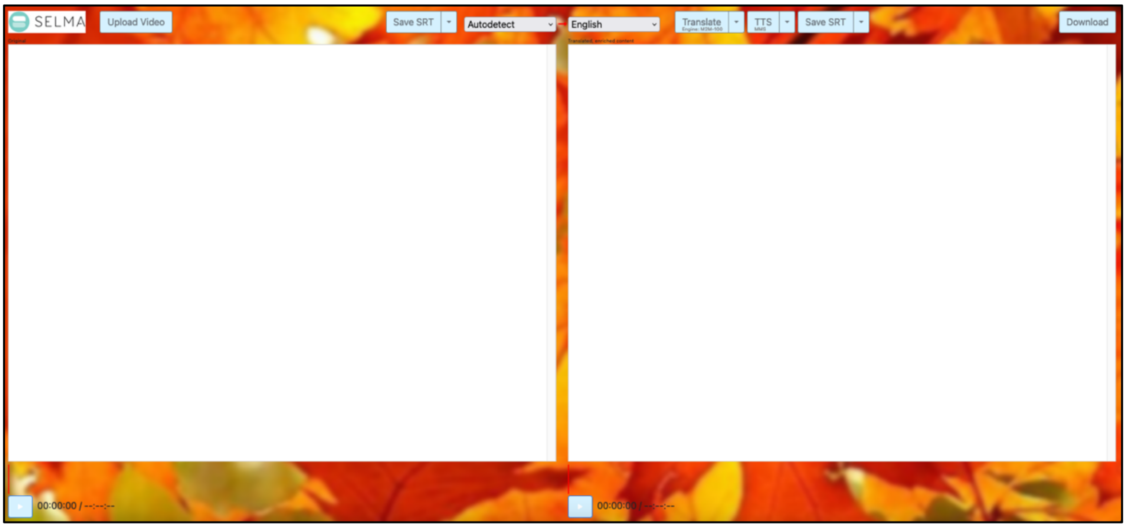
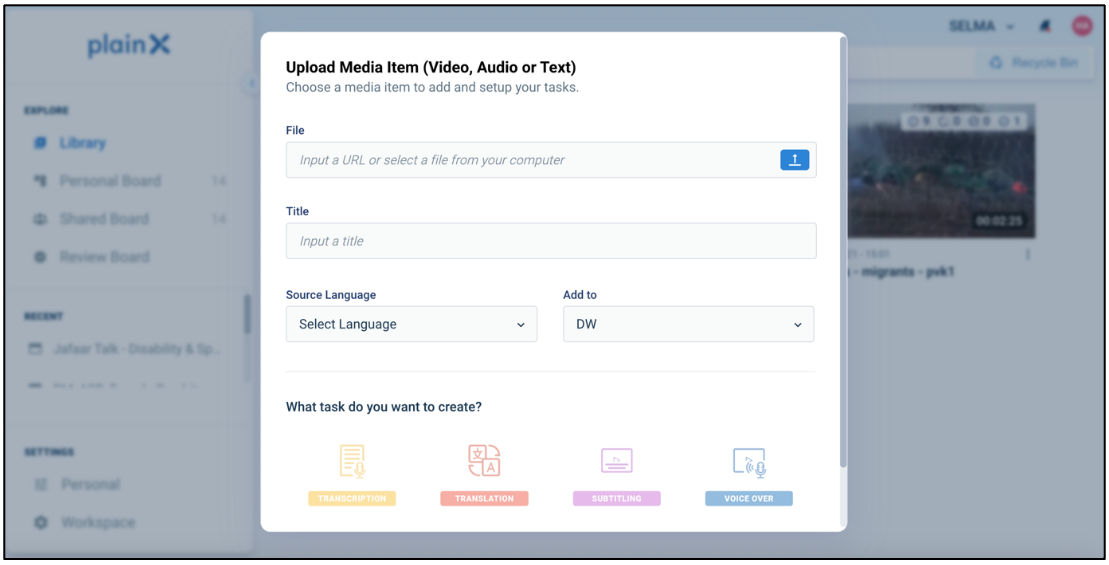


Figure 5 *SELMA OSS Interface*

Overall, the SELMA open-source platform requires some prior knowledge to be able to use it effectively but is straightforward to use and therefore easily adaptable to other use cases.

In plain X, users are guided through the addition of items or the creation of tasks. The UI clearly shows where and how to add an item and to create a task. The color coding is also clearly presents and helps users to situate themselves in the platform.



Upload Media Item (Video, Audio or Text)

Choose a media item to add and setup your tasks.

File: Input a URL or select a file from your computer

Title: Input a title

Source Language: Select Language

Add to: DW

What task do you want to create?

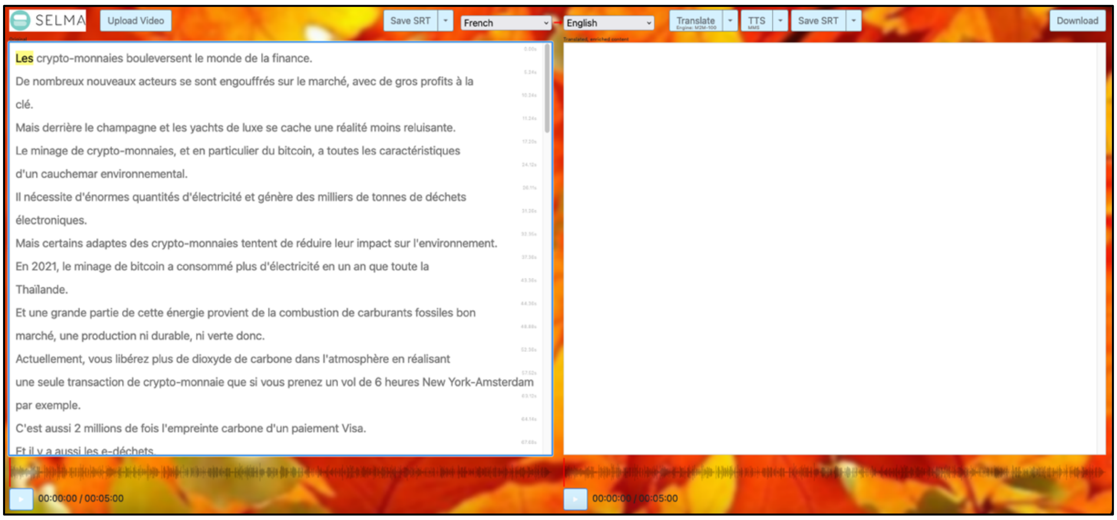
Figure 6 *SELMA OSS Interface*

This setup allows users to quickly understand how to navigate the platform and how to create tasks. However, it also limits its flexibility and its application to other use cases.

Upload of a video file

In SELMA OSS, users can upload a video and get a transcription back. A major feature is the automatic detection of the source language: users don't have to select the source language of an item manually.

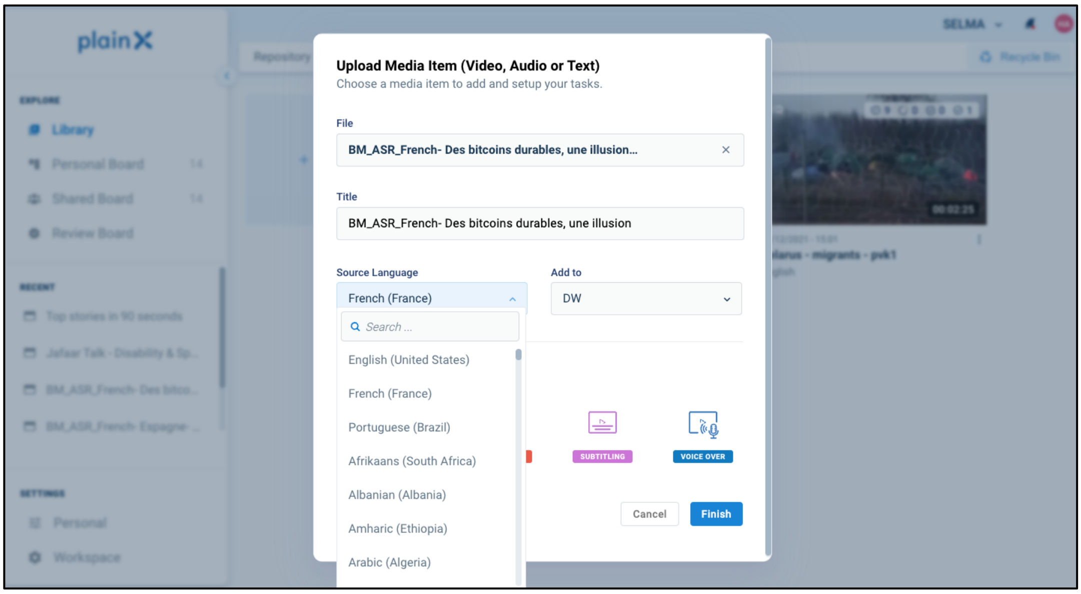
The upload and transcription are fast: for a video of 5 minutes, the upload and transcription of the video happened in less than 1 min. The user cannot choose which engine or provider to use for a transcription.



Les crypto-monnaies bouleversent le monde de la finance.
De nombreux nouveaux acteurs se sont engouffrés sur le marché, avec de gros profits à la clé.
Mais derrière le champagne et les yachts de luxe se cache une réalité moins reluisante.
Le minage de crypto-monnaies, et en particulier du bitcoin, a toutes les caractéristiques d'un cauchemar environnemental.
Il nécessite d'énormes quantités d'électricité et génère des milliers de tonnes de déchets électroniques.
Mais certains adaptes des crypto-monnaies tentent de réduire leur impact sur l'environnement.
En 2021, le minage de bitcoin a consommé plus d'électricité en un an que toute la Thaïlande.
Et une grande partie de cette énergie provient de la combustion de carburants fossiles bon marché, une production ni durable, ni verte donc.
Actuellement, vous libérez plus de dioxyde de carbone dans l'atmosphère en réalisant une seule transaction de crypto-monnaie que si vous prenez un vol de 6 heures New York-Amsterdam par exemple.
C'est aussi 2 millions de fois l'empreinte carbone d'un paiement Visa.
Et il y a aussi les e-déchets.

***Figure** 7 SELMA OSS Upload Function*

Uploading a video in plain X requires more time. A key difference between the two platforms is that in plain X, the user needs to select the source language as well as the appropriate variant.



Upload Media Item (Video, Audio or Text)

Choose a media item to add and setup your tasks.

File: BM_ASR_French- Des bitcoins durables, une illusion...

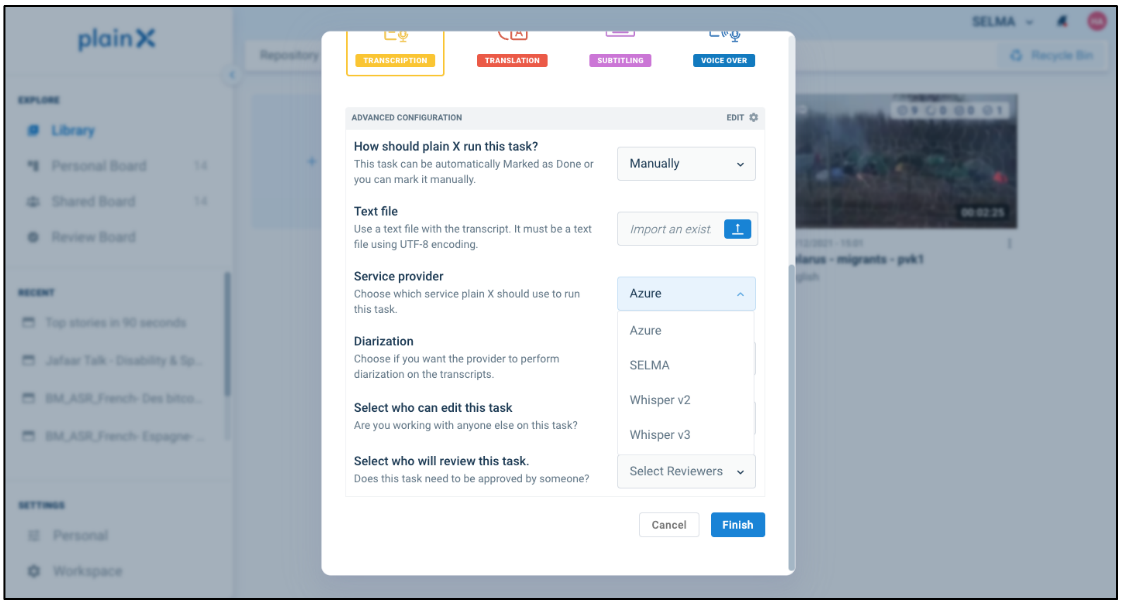
Title: BM_ASR_French- Des bitcoins durables, une illusion

Source Language: French (France)

Add to: DW

Figure 8 plain X Upload Function

They can also choose which provider to use for their transcription task and whether to activate the diarization (speaker segmentation) or not.



ADVANCED CONFIGURATION

How should plain X run this task?
This task can be automatically Marked as Done or you can mark it manually. — Manually

Text file
Use a text file with the transcript. It must be a text file using UTF-8 encoding. — Import an exist

Service provider
Choose which service plain X should use to run this task. — Azure

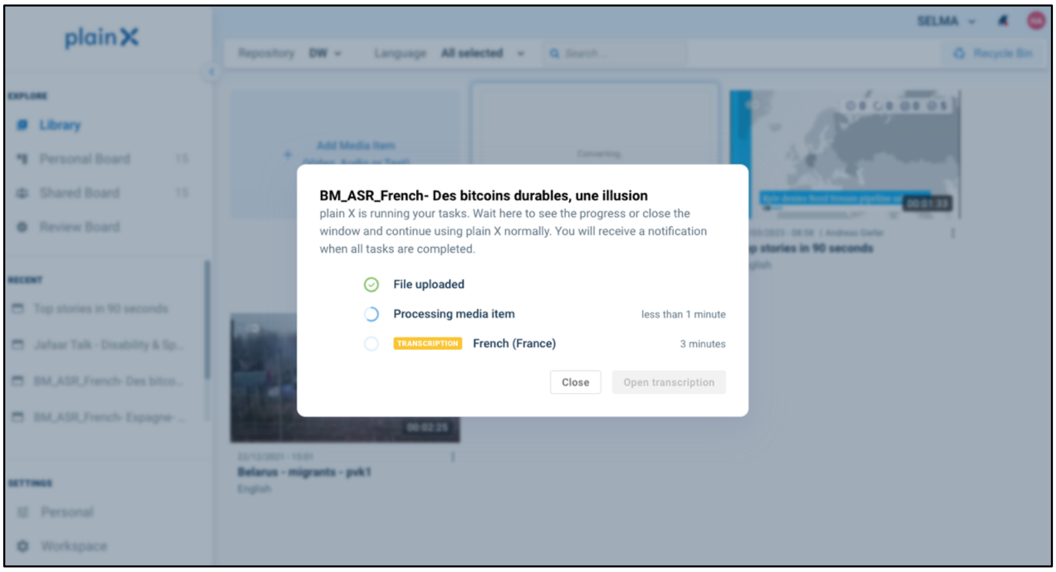
Diarization
Choose if you want the provider to perform diarization on the transcripts.

Select who can edit this task
Are you working with anyone else on this task?

Select who will review this task.
Does this task need to be approved by someone? — Select Reviewers

Figure 9 plain Transcription Options

The upload and processing of the same video item is slightly longer in plain X than in SELMA OSS. The transcription takes significantly more time on plain X.



BM_ASR_French- Des bitcoins durables, une illusion

plain X is running your tasks. Wait here to see the progress or close the window and continue using plain X normally. You will receive a notification when all tasks are completed.

File uploaded

Processing media item less than 1 minute

TRANSCRIPTION French (France) 3 minutes

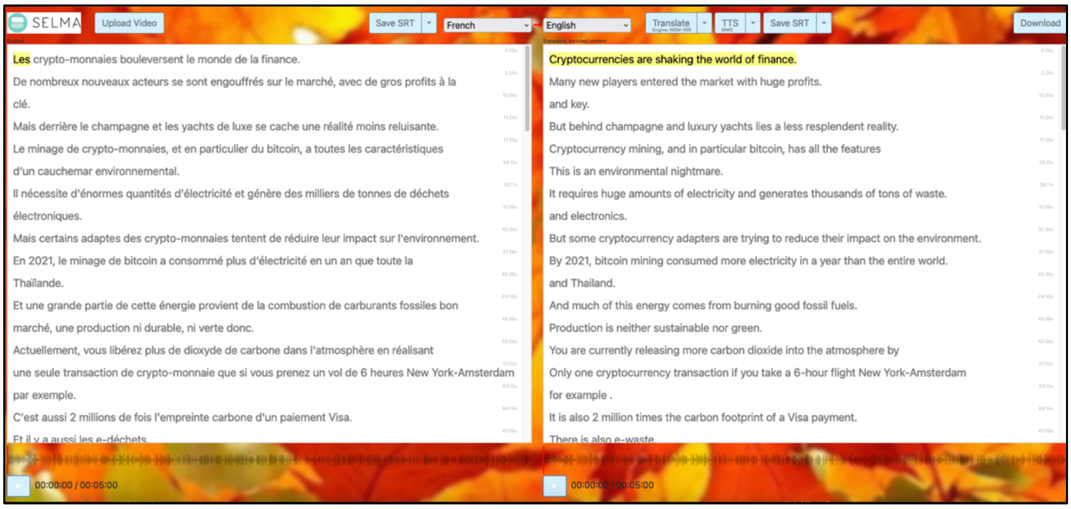
Close Open transcription

__Figure 10__ plain X Transcription Time Needed

Translation

SELMA OSS and plain X both offer translation as a feature. In both platforms, users can choose which engine to use. The choice is larger with plain X, where commercial providers are also available. For example, for a translation from French to English, plain X offers 4 different providers (Azure, Deepl, Google, Facebook) whereas SELMA OSS gives the choice between two open-source models (M2M-100 from Facebook or HuggingFace models).

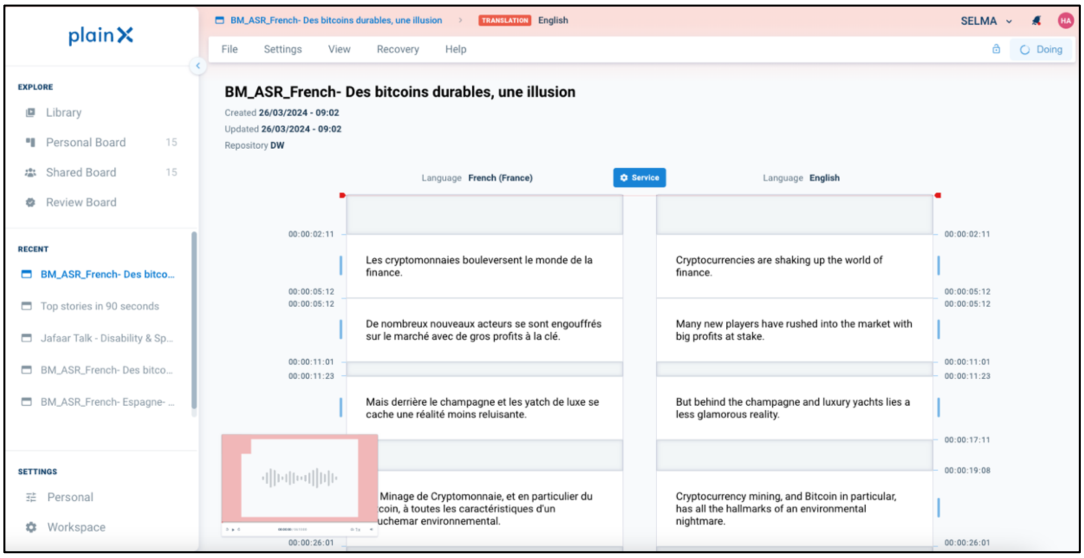
In terms of speed, both platforms are similar and provide a translation from text in less than a minute.



Les crypto-monnaies bouleversent le monde de la finance.
De nombreux nouveaux acteurs se sont engouffrés sur le marché, avec de gros profits à la
clé.
Mais derrière le champagne et les yachts de luxe se cache une réalité moins reluisante.
Le minage de crypto-monnaies, et en particulier du bitcoin, a toutes les caractéristiques
d'un cauchemar environnemental.
Il nécessite d'énormes quantités d'électricité et génère des milliers de tonnes de déchets
électroniques.
Mais certains adaptes des crypto-monnaies tentent de réduire leur impact sur l'environnement.
En 2021, le minage de bitcoin a consommé plus d'électricité en un an que toute la
Thaïlande.
Et une grande partie de cette énergie provient de la combustion de carburants fossiles bon
marché, une production ni durable, ni verte donc.
Actuellement, vous libérez plus de dioxyde de carbone dans l'atmosphère en réalisant
une seule transaction de crypto-monnaie que si vous prenez un vol de 6 heures New York-Amsterdam
par exemple.
C'est aussi 2 millions de fois l'empreinte carbone d'un paiement Visa.
Et il y a aussi les e-déchets

Cryptocurrencies are shaking the world of finance.
Many new players entered the market with huge profits.
and key.
But behind champagne and luxury yachts lies a less resplendent reality.
Cryptocurrency mining, and in particular bitcoin, has all the features
This is an environmental nightmare.
It requires huge amounts of electricity and generates thousands of tons of waste.
and electronics.
But some cryptocurrency adapters are trying to reduce their impact on the environment.
By 2021, bitcoin mining consumed more electricity in a year than the entire world.
and Thailand.
And much of this energy comes from burning good fossil fuels.
Production is neither sustainable nor green.
You are currently releasing more carbon dioxide into the atmosphere by
Only one cryptocurrency transaction if you take a 6-hour flight New York-Amsterdam
for example .
It is also 2 million times the carbon footprint of a Visa payment.
There is also e-waste

***Figure 11** SELMA OSS Translation*


plainX
BM_ASR_French- Des bitcoins durables, une illusion
TRANSLATION
English
SELMA
HA
File
Settings
View
Recovery
Help
Doing
EXPLORE
Library
Personal Board 15
Shared Board 15
Review Board
RECENT
BM_ASR_French- Des bitco...
Top stories in 90 seconds
Jafaar Talk - Disability & Sp...
BM_ASR_French- Des bitco...
BM_ASR_French- Espagne- ...
SETTINGS
Personal
Workspace
Service

BM_ASR_French- Des bitcoins durables, une illusion
Created 26/03/2024 - 09:02
Updated 26/03/2024 - 09:02
Repository DW

Language French (France)
Language English

00:00:02:11 00:00:05:12	Les cryptomonnaies bouleversent le monde de la finance.	Cryptocurrencies are shaking up the world of finance.	00:00:02:11 00:00:05:12
00:00:05:12 00:00:11:01	De nombreux nouveaux acteurs se sont engouffrés sur le marché avec de gros profits à la clé.	Many new players have rushed into the market with big profits at stake.	00:00:05:12 00:00:11:01
00:00:11:23	Mais derrière le champagne et les yatch de luxe se cache une réalité moins reluisante.	But behind the champagne and luxury yachts lies a less glamorous reality.	00:00:11:23 00:00:17:11
00:00:26:01	Minage de Cryptomonnaie, et en particulier du coin, à toutes les caractéristiques d'un uchemar environnemental.	Cryptocurrency mining, and Bitcoin in particular, has all the hallmarks of an environmental nightmare.	00:00:19:08 00:00:26:01

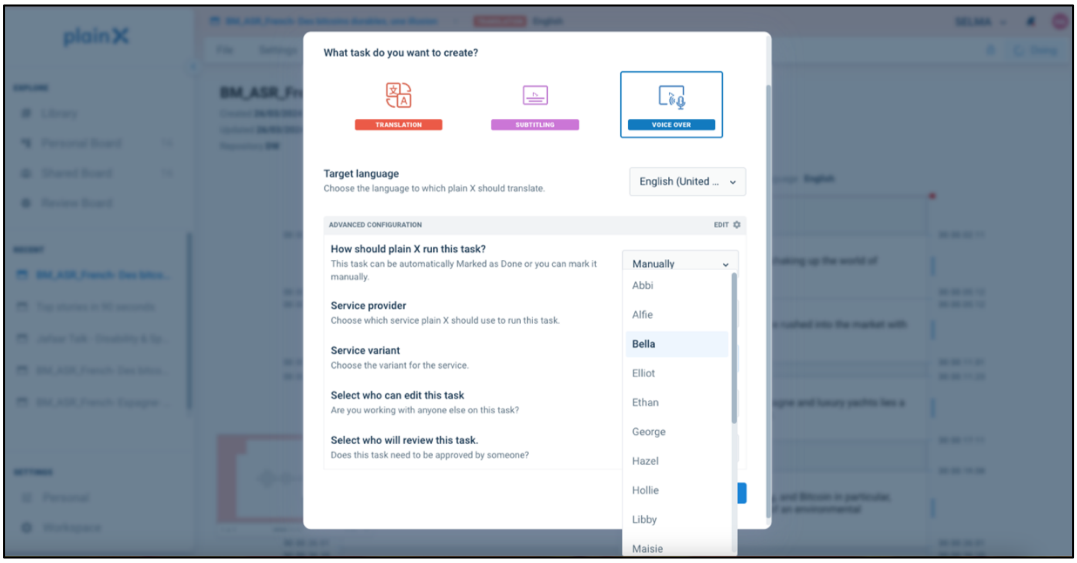
***Figure 12** plain X Translation*

Voice-Over

Both platforms can create voice-overs from a translated text.

In SELMA OSS, users can create a voice-over by using the "TTS" button. They do not have a choice in which provider or synthetic voice to use.

In plain X, users can choose from different providers and synthetic voices.



What task do you want to create?

Target language
Choose the language to which plain X should translate.
English (United ...

ADVANCED CONFIGURATION
EDIT

How should plain X run this task?
This task can be automatically Marked as Done or you can mark it manually.
Manually

Service provider
Choose which service plain X should use to run this task.

Service variant
Choose the variant for the service.

Select who can edit this task
Are you working with anyone else on this task?

Select who will review this task.
Does this task need to be approved by someone?

Abbi
Alfie
Bella
Elliot
Ethan
George
Hazel
Hollie
Libby
Maisie

Figure 13 *plain X Voice-Over Functionality*

For this task, plain X performs significantly faster. The voice-over creation for a 5-minute video took less than a minute. In SELMA OSS, for the same video, the creation of a voice-over took double the time.

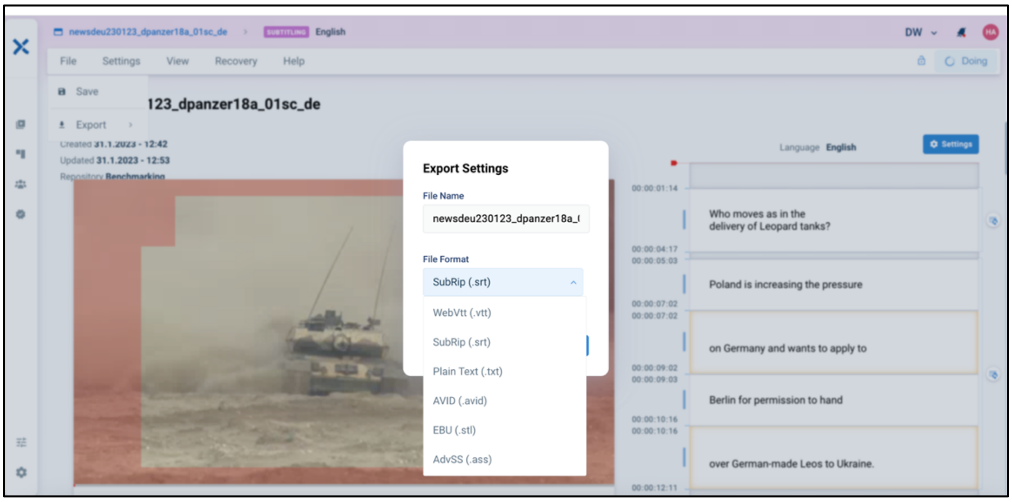
Export Options

In SELMA OSS, the transcription and the translation can only be downloaded in a SubRip (.srt) format. The voice-over is downloaded in a WAVE (.wav) format.

For each task in plain X, users can decide in which format to export the task. The different formats available in plain X are:

- Transcription: Plain Text (.txt), SubRip (.srt)
- Translation: Plain Text (.txt)
- Subtitles: WebVtt (.vtt), SubRip(.srt), Plain Text (.txt), AVID (.avid), EBU STL (.stl), AdvSS (.ass)

Overall, the SELMA OSS output is more suitable for integration within other platforms. For example, the translation output is integrated in the Benchmarking tool. In plain X, users have the choice to use formats that they are more familiar with. Standard formats can also be more easily integrated with 3rd party applications.



newsdeu230123_dpanzer18a_01sc_de › SUBTITLING English

123_dpanzer18a_01sc_de

Created 31.1.2023 - 12:42
Updated 31.1.2023 - 12:53
Repository Benchmarking

Export Settings

File Name: newsdeu230123_dpanzer18a_(

File Format: SubRip (.srt)

Language English

00:00:01:14
Who moves as in the
delivery of Leopard tanks?

00:00:04:17
00:00:05:03
Poland is increasing the pressure

00:00:07:02
00:00:07:02
on Germany and wants to apply to

00:00:09:02
00:00:09:03
Berlin for permission to hand

00:00:10:16
00:00:10:16
over German-made Leos to Ukraine.

00:00:12:11

***Figure 14** Export formats in plain X*

As to the quality output and usefulness, we have to take into consideration that the SELMA OSS is a public platform and costs need to be minimized. Using expensive engines would drive up costs very fast. Therefore, we must keep in mind that free, open-source engines are used, e.g. HuggingFace or M2M-100 from Facebook for MT, with a quality that is generally not comparable with commercial services. Whisper is used for ASR, and transcription quality differs widely according to language.

MT and ASR quality assessment is part of our benchmarking effort, and the quality output would be similar for one and the same engine, regardless of the platform, whether we have processed it through plain X or through SELMA OSS. We confirmed this with a benchmarking assessment. Of course, a more advanced platform like plain X offers more providers with high-quality engines, resulting in a higher output quality.

Accessing the OSS via API was used by the DW Speaker and Podcast Creator applications. More specifically, the OSS platform hosted the services dedicated to converting text into synthetic speech in Brazilian Portuguese and Urdu. First experiments with the API revealed instabilities in the system that could later be reduced to a minimum. Once stabilized, the API showed itself to be very responsive, returning the generated audio files after an acceptable processing time. As an example, it took 17 seconds to generate the speech for a 3-minute long news bulletin in Urdu.

We can **conclude** that the OSS serves the user that is looking for a very simple, easy to access, non-technical, tool for transcription, translation or voice-over, especially if that is an online system. As long as there is a (free) online service, that is easiest for the non-expert user. However, this is not foreseen as a long-term objective of the SELMA project. If it is not available online, they will have to download it and install it on their computer, which many users will refrain from doing.

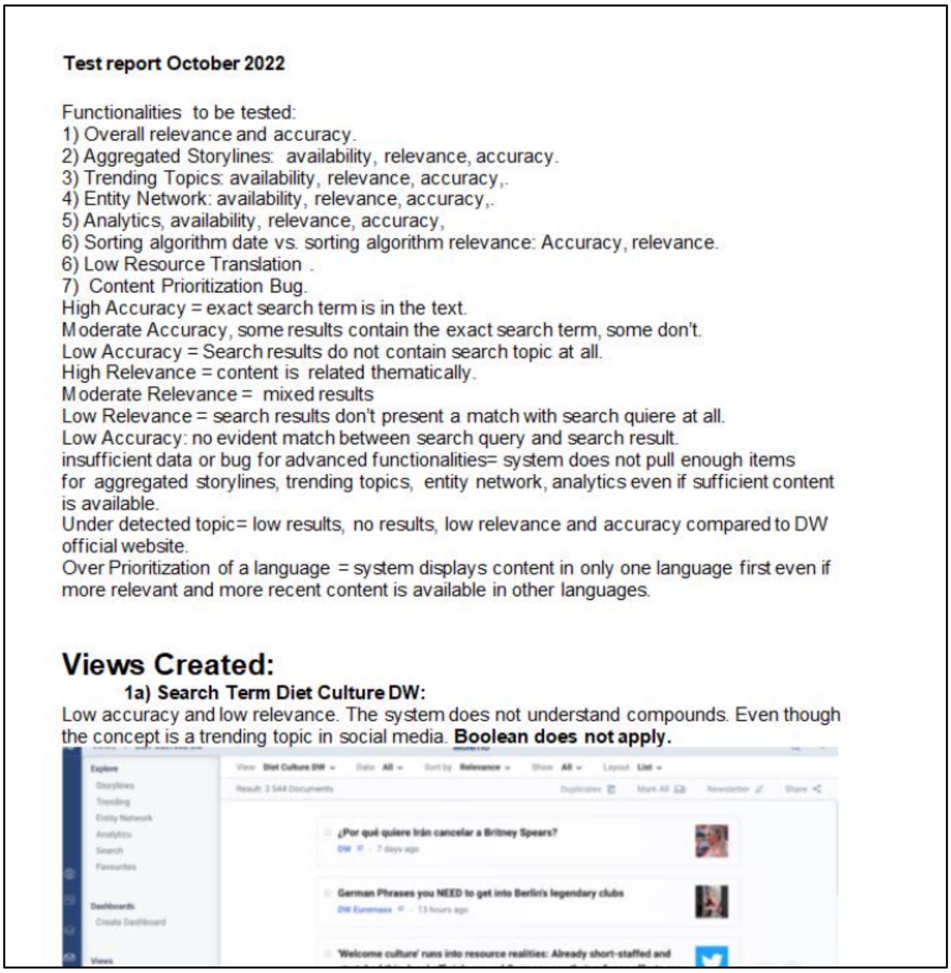
For SMEs and user groups that are not specialized in language technologies or require advanced and high-volume NLP processing, this open-source very simple software kit may be the low-cost and simple solution for ad-hoc processing. It is flexible, as other engines can be connected to the system and it can run on a simple CPU. It may also serve as a first acquaintance with such a language processing workflow and users may upgrade to a fully professional system in time.

4.2 Monitio Demonstrator

The Monitio platform is the main demonstrator for the first use case (UC1), multilingual media monitoring, as described in D1.1 - Use Case Description and Requirements. It scans a wide collection of media items and provides a sophisticated filtering and automatic analysis system, producing a selection of clustered news items by topic or another common attribute, according to the user's preferences.

In **Year 1**, the platform was introduced to the consortium and requirements and user scenarios were set. It was decided which NLP enhancements to the tool were needed and were within the scope of the project and the consortium. In particular improvements as to integration, analysis speed, named entity recognition and linking, building dictionaries and integrating thesauri, incorporating user feedback mechanisms, and user interfacing were addressed.

Throughout **Year 2**, Deutsche Welle tried out the Monitio platform and provided feedback on the overall UI, functionalities, ease of use, transparency and consistency. We discussed possible use cases internally with media professionals, setting priorities and suggesting changes. Content feeds were provided by Deutsche Welle and integrated into the platform. After numerous trials, it was decided that the newsletter production is a prime goal for wider distribution and translation of the headlines and cluster summarizations into the user's preferred target language. Detailed input on functionalities such as filtering, search parameters, language settings, was given. The focus on text analysis makes sense, but the content ingestion should be expanded to include video and audio formats. YouTube videos were added in this reporting period on user demand. A renewed UI was developed towards the end of the year and testing is currently underway, with regular feedback.



Test report October 2022

Functionalities to be tested:
1) Overall relevance and accuracy.
2) Aggregated Storylines: availability, relevance, accuracy.
3) Trending Topics: availability, relevance, accuracy,.
4) Entity Network: availability, relevance, accuracy,.
5) Analytics, availability, relevance, accuracy,
6) Sorting algorithm date vs. sorting algorithm relevance: Accuracy, relevance.
6) Low Resource Translation .
7) Content Prioritization Bug.
High Accuracy = exact search term is in the text.
Moderate Accuracy, some results contain the exact search term, some don't.
Low Accuracy = Search results do not contain search topic at all.
High Relevance = content is related thematically.
Moderate Relevance = mixed results
Low Relevance = search results don't present a match with search quiere at all.
Low Accuracy: no evident match between search query and search result.
insufficient data or bug for advanced functionalities= system does not pull enough items for aggregated storylines, trending topics, entity network, analytics even if sufficient content is available.
Under detected topic= low results, no results, low relevance and accuracy compared to DW official website.
Over Prioritization of a language = system displays content in only one language first even if more relevant and more recent content is available in other languages.

Views Created:

1a) Search Term Diet Culture DW:

Low accuracy and low relevance. The system does not understand compounds. Even though the concept is a trending topic in social media. **Boolean does not apply.**

View Diet Culture DW · Date All · Sort by Relevance · Show All · Layout List

Result: 3 544 Documents

¿Por qué quiere Irán cancelar a Britney Spears?
DW · 7 days ago

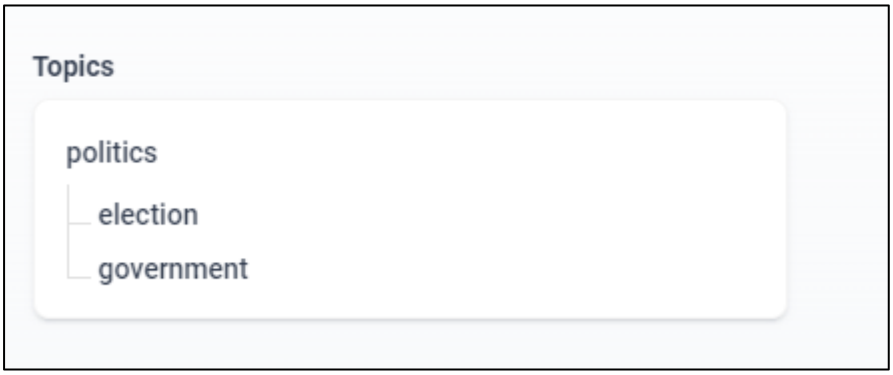
German Phrases you NEED to get into Berlin's legendary clubs
DW Euromaxx · 13 hours ago

'Welcome culture' runs into resource realities: Already short-staffed and

Figure 15 *Example of feedback on the Monitio functionalities*

Year 3 included a thorough and wide user evaluation of the updated/finalized platform, with different departments, including DW Archive and Documentation for specialized searching, retrieval and integration, and editorial language departments for internal monitoring of DW content and creation and use of customized Monitio Newsletters.

NER and topics were enhanced. We now have a hierarchical presentation of the topics, which was a key user demand, as it makes the use of topic labeling much more powerful. This opens up the possibility of eventually integrating the hierarchical internal topical database that DW uses. It is essential to have a consistent, structured, and wide-ranging set of topics and keywords, to arrive at effective searching and clustering.



Topics

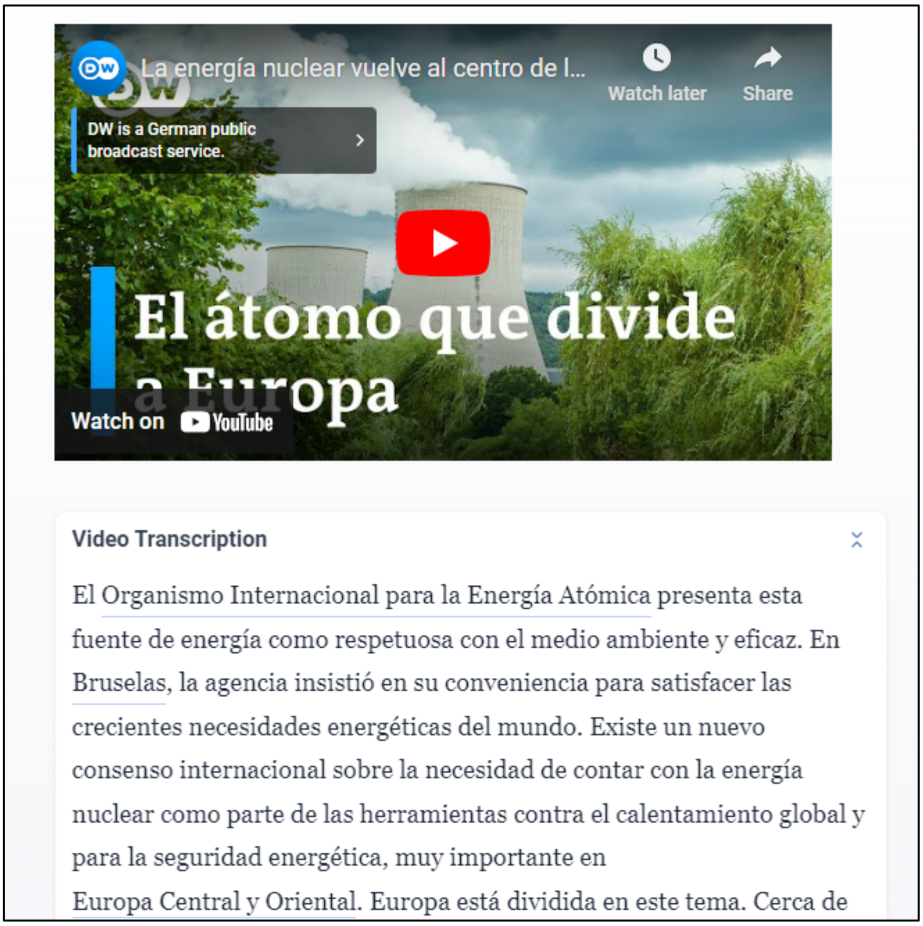
politics

election

government

Figure 16 *Hierarchical topics list as requested by users*

Also the annotation of named entities in the article text or the transcribed text is considered very helpful, as we can easily scan the content that way. It would be good if that annotation can also be applied to the translation, as now it is lost after translation.



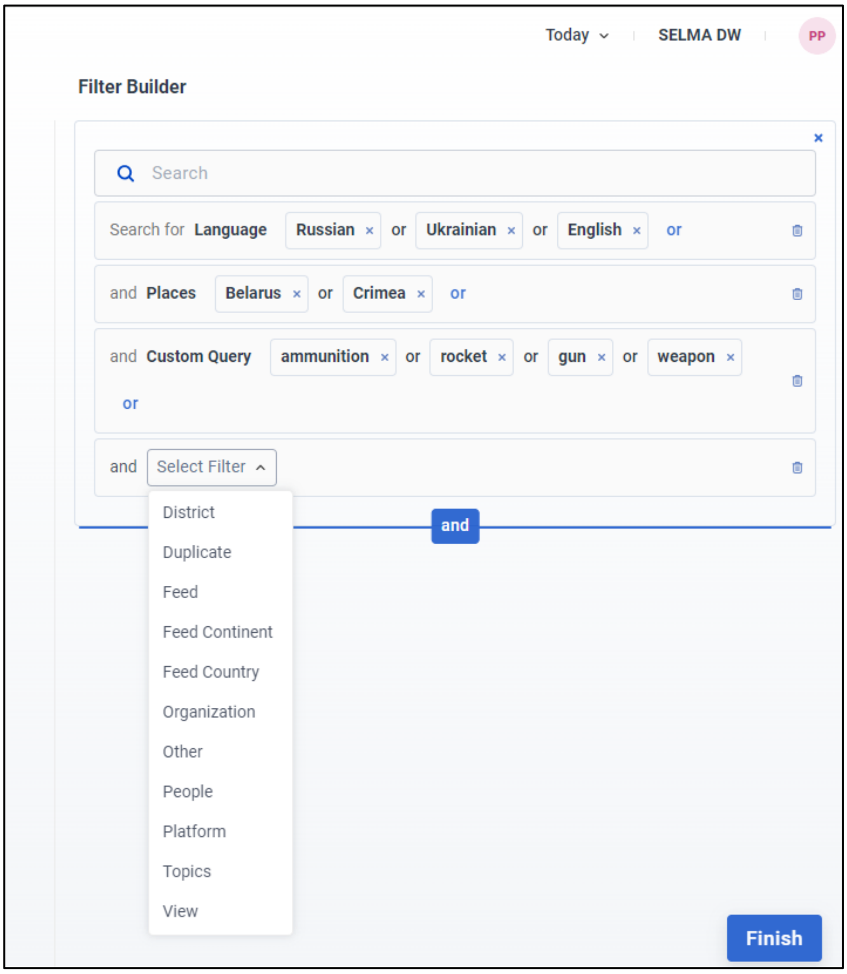
La energía nuclear vuelve al centro de l...
Watch later
Share
DW is a German public broadcast service.
El átomo que divide a Europa
Watch on YouTube
Video Transcription

El Organismo Internacional para la Energía Atómica presenta esta fuente de energía como respetuosa con el medio ambiente y eficaz. En Bruselas, la agencia insistió en su conveniencia para satisfacer las crecientes necesidades energéticas del mundo. Existe un nuevo consenso internacional sobre la necesidad de contar con la energía nuclear como parte de las herramientas contra el calentamiento global y para la seguridad energética, muy importante en Europa Central y Oriental. Europa está dividida en este tema. Cerca de

Figure 17 *Named Entity annotation in the text is very helpful*

The Monitio NER analysis tool was made available via API and the DW team uses it for NER benchmarking in ASR output (see the section on Benchmarking). It is also being considered for use to train the English live subtitling system currently in use at Deutsche Welle and this NER tool could provide regular (e.g. weekly) regular input to train the ASR engine.

It is considered a great improvement that we can now set up an advanced, customized **search strategy**, which we can save and share and run on a regular basis. We can put that in a standard report, add an interactive chart, and generate and publish it as a regularly occurring newsletter distributed to certain departments or user groups. It was essential that the search strategy had all the functions of advanced searching, including free text used as keywords. This was the kind of feedback loop from the user partner to the developers that has resulted in an enhanced and usable tool. As a result, we can filter on any language, date, source, and add a series of topics or free-text keywords with Boolean AND or OR. This is now also presented in a transparent, user-friendly way. Initially, it was not possible, then it was enabled with a somewhat obscure trick, and, finally, well integrated in the UI, clear to the user.



Filter Builder

Search for Language Russian or Ukrainian or English or

and Places Belarus or Crimea or

and Custom Query ammunition or rocket or gun or weapon or

and Select Filter

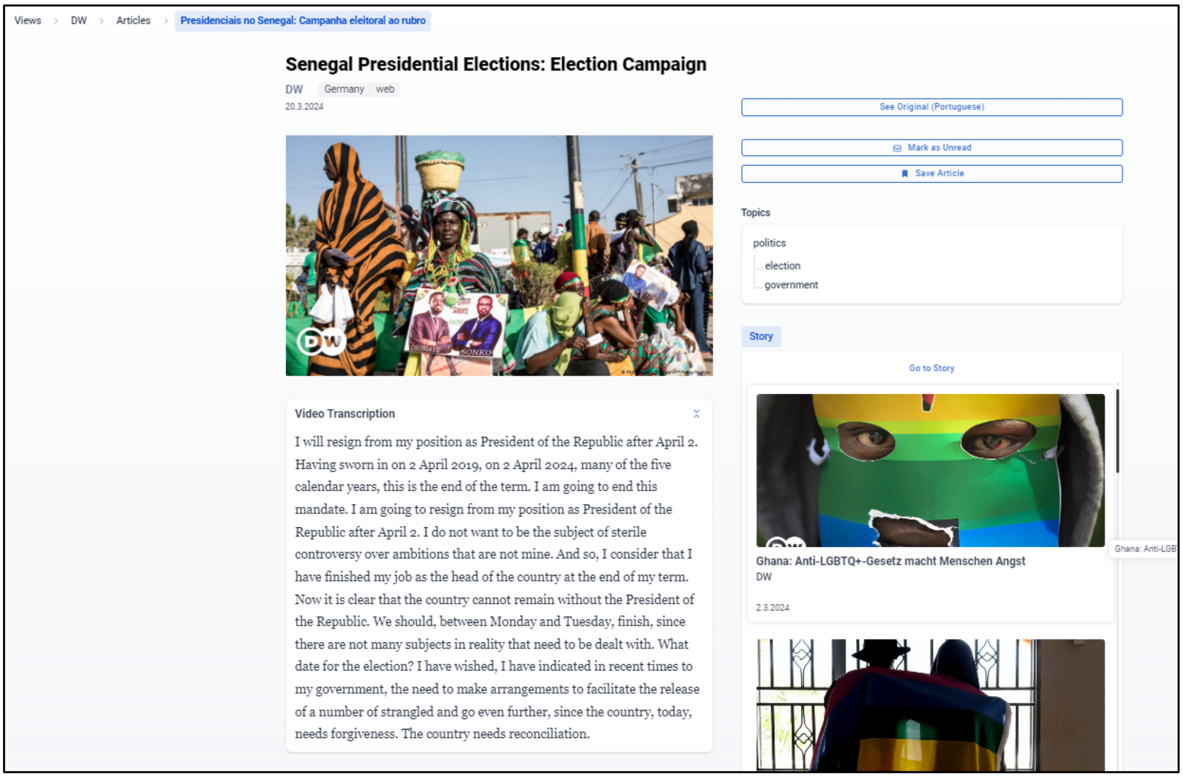
***Figure 18** Customized advanced saved search strategy*

The platform was evaluated by several user groups at Deutsche Welle. The first group were specific editorial departments, e.g. Hindi and Urdu language departments, who assessed it in terms of usability, ease of use and further requirements. Their interest was mostly on the use of the platform for **external monitoring**: a daily digest of what was published in the media (including websites, newspapers, social media, etc) in their regional coverage, for instance in India. A search strategy can then be set up, saved, run and displayed at the daily briefing to determine what articles should be produced that day/the next few days. Their assessment of the tool was that it was easy to use, clear, and can be used immediately if the external sources that they need can be added to the platform. This request – and a list of required sources – has been forwarded to Priberam. The issue is licenses for monitoring such sources. A solution that has been discussed is to restrict access to the content by making it only available as a link to the original source, which would work for this use case. A transcription and translation could be made searchable and readable in full text. Translation in this use case is not essential, as native speakers would do this kind of monitoring.

Another user group provided extensive feedback and a set of requirements also in the area of external monitoring, but for different purposes. They were looking for certain trends or biased content in particular sources for trustworthiness or finding/excluding (un)reliable content. This led to the request to enrich the saved search strategy option and make it more advanced and flexible. This search strategy function is now considered adequate. Also here, the issue of external sources needs to be solved before Monitio can be used for such use case for real.

The SELMA podcasting app (see details later) also extracts content from Monitio to compile a list of news texts based on a saved strategy. This is an example of automated media monitoring and the result of extensive trial and error, with ultimately a very nice outcome.

The third user group looked at **internal monitoring**, analyzing DW content that is present in the database. The first requirement that was passed to the developer, Priberam, was that we need to cover virtually all of DW content, in all formats, from all channels and in all languages to make this into an efficient internal monitoring tool. Especially transcription of videos and audios was essential. Initially, only YouTube could be transcribed, with a considerable delay in time. Content on DW's regular website (www.dw.com) only included text articles in full and metadata without transcript for videos. That was not sufficient. Gradually, more sources, channels and languages were added, videos that were embedded were included, and all video was transcribed and translated into English. This makes the platform an extremely powerful tool in which all DW content, and over several channels and sources, in any language can be monitored. Translation into English is essential here, as the monitoring users obviously do not master all DW covered languages (currently 32).


Views › DW › Articles › Presidenciais no Senegal: Campanha eleitoral ao rubro
See Original (Portuguese)
Mark as Unread
Save Article
Topics
politics
election
government
Story
Go to Story
Ghana: Anti-LGBTQ+-Gesetz macht Menschen Angst
DW
2.3.2024
Ghana: Anti-LGB

Senegal Presidential Elections: Election Campaign

DW Germany | web

20.3.2024

Video Transcription

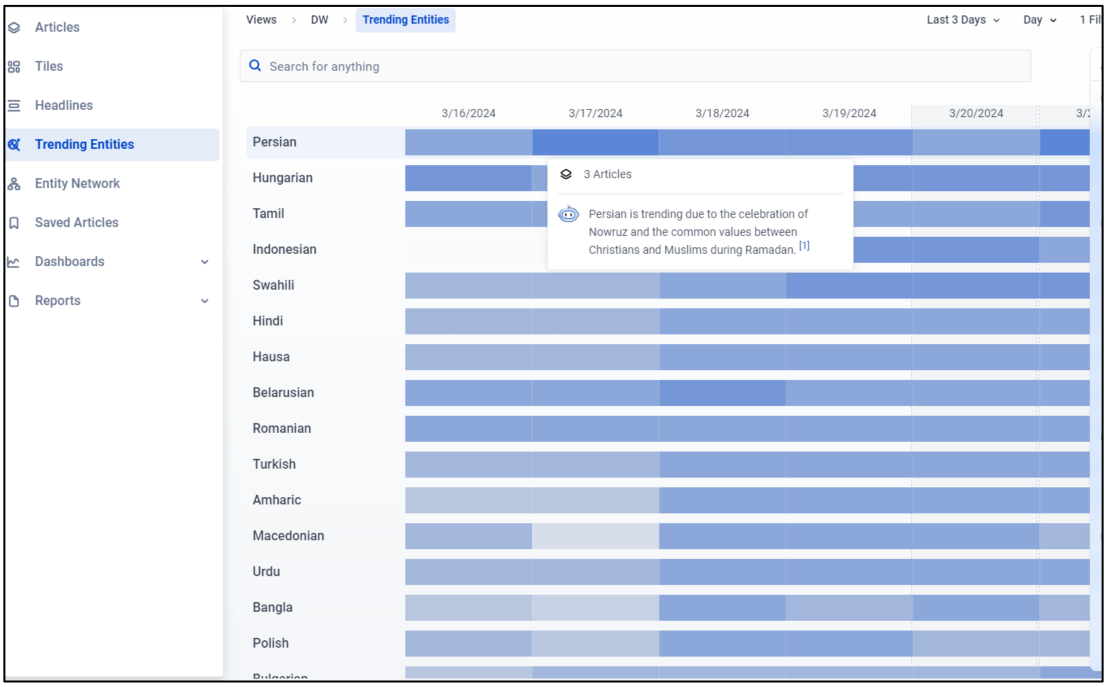
I will resign from my position as President of the Republic after April 2. Having sworn in on 2 April 2019, on 2 April 2024, many of the five calendar years, this is the end of the term. I am going to end this mandate. I am going to resign from my position as President of the Republic after April 2. I do not want to be the subject of sterile controversy over ambitions that are not mine. And so, I consider that I have finished my job as the head of the country at the end of my term. Now it is clear that the country cannot remain without the President of the Republic. We should, between Monday and Tuesday, finish, since there are not many subjects in reality that need to be dealt with. What date for the election? I have wished, I have indicated in recent times to my government, the need to make arrangements to facilitate the release of a number of strangled and go even further, since the country, today, needs forgiveness. The country needs reconciliation.

Figure 19 *Users can now get translated video transcription for all languages*

This user group included archivists and staff from the strategy departments and overall management. Currently, YouTube and dw.com content is covered. Additional channels such as Facebook, TikTok and Instagram would complete the picture.

There are many use cases for this kind of internal monitoring, for instance, management getting a daily overview of what was published the day/week before, archive running searches on specific topics over the past year, journalists looking for content on a certain topic from other language departments for research or reuse (this is quite hard to get at the moment). The combination of transcription and translation of audio and video content, allowing analysis, search and retrieval over historic content in all DW languages opens up many doors to efficient monitoring.

The **Trending Entities** module offers a very useful feature for DW that was not initially foreseen by the developers. In addition to getting a list of trending topical keywords or named entities, a filtering on language only provides a very useful and up to date overview of the volume or **productivity** of DW language departments over time, as well as an indication of the most trending topics in that language.



Persian is trending due to the celebration of Nowruz and the common values between Christians and Muslims during Ramadan. [1]

__Figure 20__ Overview of productivity per language using Trending Entities

The enhanced **Clustering** feature was tested throughout the coverage period, as it is part of the general search and demo activities. It offers a solution to a specific user demand to find duplicate (or near duplicate) content in one language or over several languages. This allows us to identify items that have been published on different channels and thus determine the productivity and reuse of the content. That is also why it is important to include as many distribution channels as possible (including social media channels). This is now possible with the “duplicate” and ”disaggregate duplicates” buttons. An extended request in this regard is to expand this to locate duplicates over different languages, so that we can find which items have been published in different language channels. Similarly, this function is also used the other way around: to avoid duplicates, so a search result does not come up with different versions of the same item.

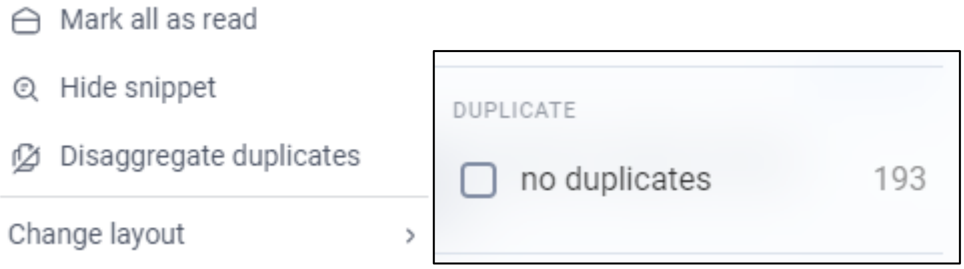


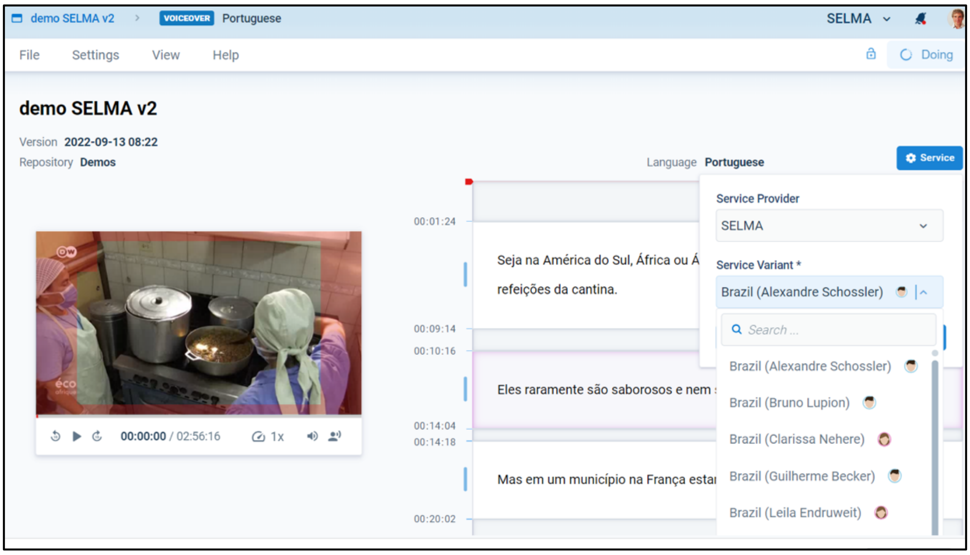
Figure 21 *Users can find (near-)duplicates or exclude duplicate content*

Conclusion: Throughout the project time and especially during the final year, several user groups at DW – and some other groups, such as the German public broadcast association and the EBU, as well as potential clients in Spain and Portugal – participated in demos or trials of Monitio. It was usually met with great appreciation, and users find it a very powerful platform, especially because of its multilinguality, advanced analysis and clustering functioning, which they have not seen in other platforms. It has turned into a tool with many opportunities and suitable for different use cases. It has demonstrated the benefits of analyzing an organization's own content. If we can solve the issue of licenses/right restrictions for external sources and include those, also the external monitoring is off to a good start.

4.3 plain X Demonstrator

In **Year 1**, we established the requirements, scenarios and evaluation plan for the content creation use case (UC2), with the plain X demonstrator as the prime target platform. In this period, we determined the initial status of the newsbridge/plain X platform, which has been developed over several years and is a platform which is being further jointly developed and exploited by Priberam and DW. It targets a smooth workflow for multilingual transcription, translation, subtitling and voice-over with synthetic voices. User requirements, workflows and enhancements that can be developed within SELMA were discussed.

Year 2 focused on evaluating envisaged integrated enhancements, including ASR modules, speech-to-translated-text and customized synthetic voices for Brazilian Portuguese, all modules from the University of Avignon. End users started evaluating the output and comparing it with other processes and tools. This is described in D5.2, sections 4.7, 4.8, and 4.9.



demo SELMA v2

Version 2022-09-13 08:22

Repository Demos

Language Portuguese

Service Provider: SELMA

Service Variant *: Brazil (Alexandre Schossler)

- Brazil (Alexandre Schossler)
- Brazil (Bruno Lupion)
- Brazil (Clarissa Nehere)
- Brazil (Guilherme Becker)
- Brazil (Leila Endruweit)

00:01:24
Seja na América do Sul, África ou Á
refeições da cantina.
00:09:14

00:10:16
Eles raramente são saborosos e nem
00:14:04

00:14:18
Mas em um município na França estai
00:20:02

Figure 22 *Brazilian customized voices in plain X*

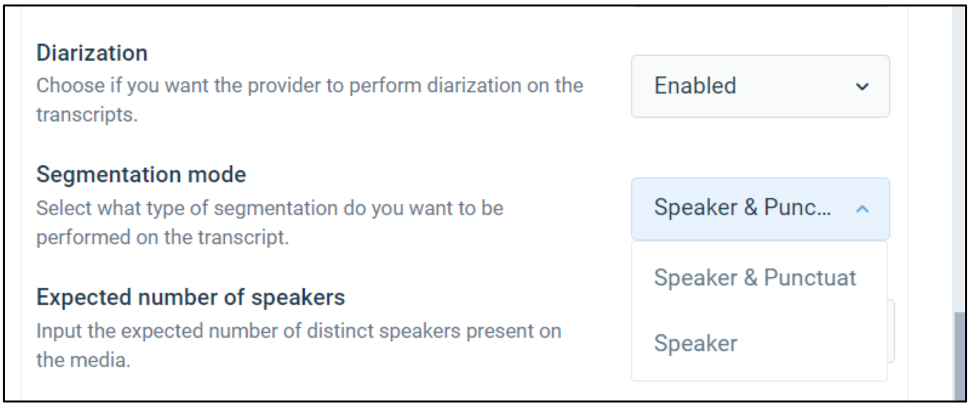
In **Year 3**, we focused on overall usability as well as new features that were developed or became available for evaluation.

As the platform matured and more users started working with the system, new requirements came up. Of course, not every user demand can be met or is even justified or appropriate. Each request or suggestion was considered, discussed and, if accepted, added as a ticket.

Through the continuous testing of the platform by DW's editorial departments – as well as some other clients - and the collected feedback, major feature requests emerged during this project, resulting in the following enhancements:

4.3.1 *Diarization*

Users that specialize in transcription expressed the wish to add diarization to the transcription tool. Especially those that worked previously with Amberscript said they missed that function. Therefore, we added diarization as a function in plain X, requiring an important re-work of the UI. In first instance, this is only available for those providers that have that option in their transcription tool, such as Amberscript and Azure. The plan is to have this expanded to other providers/engines as well. In the meantime, users have the option of getting speaker labels if they select the engines that include this, and we were able to try out and use this function in production and get further user feedback.



Diarization
Choose if you want the provider to perform diarization on the transcripts.
Enabled

Segmentation mode
Select what type of segmentation do you want to be performed on the transcript.
Speaker & Punc...
Speaker & Punctuat
Speaker

Expected number of speakers
Input the expected number of distinct speakers present on the media.

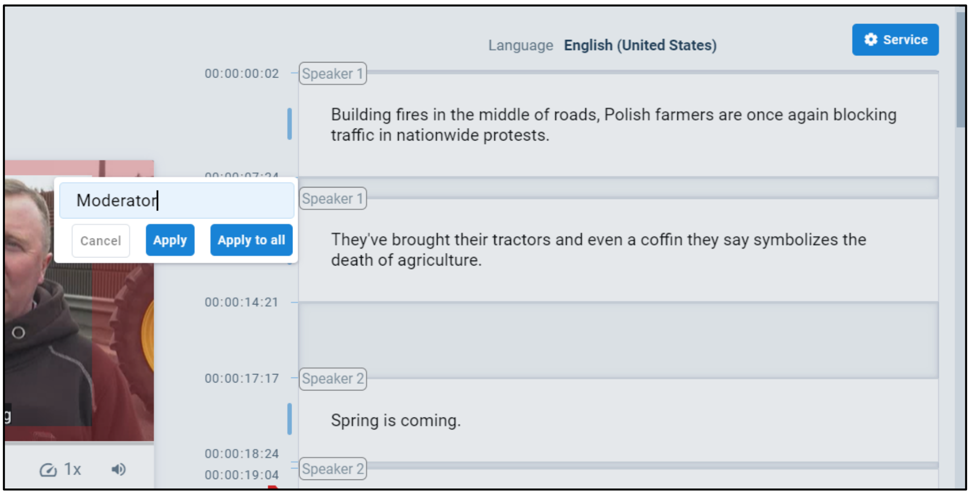
Figure 23 *Enabling diarization when creating the task*

When creating the task, the user can indicate whether they want segmentation only based on speaker labels or speaker labels combined with punctuation. This was also a specific user request, as the option with only speaker labels presents the text in paragraphs per speaker. This is used when preparing a transcribed text (possibly with subsequent translation) for (re)voicing. The text is then neatly arranged per speaker occurrence. This is also a useful format for interviews or panel discussions. The other option, i.e., speaker & punctuation, is most suited for subtitling.

When selecting diarization during the transcription creation task, the system asks for the expected number of speakers. It seemed odd initially that the speaker has to indicate this instead of the system detecting it, but some of our editorial users, in particular those very familiar with using Amberscript, confirmed that this is common procedure and this is not an obstacle for using it. Thus, the user indicates whether there are 2 or 20 speakers, for instance, and the engine uses that input as a basis for the diarization. It can, however, adapt this if more or fewer voices are detected. But an approximate number is needed to get a good labeling output.

As for the accuracy of the speaker labeling, we performed some comparative tests on how well the two engines that do diarization in plain X recognize the speaker occurrence and shift. We did this for English and German. Amberscript seems to perform slightly better in those languages, but, overall, both are fairly comparable. A correct indication of expected speakers is vital for both systems, however. Otherwise, the output shows speaker switches while the same voice continues speaking or two or more speakers get the same label.

The diarization function also allows the user to customize the labels themselves, so the initial labels of Speaker 1, Speaker 2, etc. can be replaced by the names or a description of the speaker. This can be a one-time change, for one segment, or for all segments in the entire document, so that consistent naming is applied in an efficient manner. Adding names or descriptions is also important during the transcription editing process, as it is easier to track a recognizable label than a very generic one, especially when editing a long-form video or audio with many speakers, Otherwise, the editor easily gets lost in tracking speakers.



Language English (United States)

00:00:00:02 Speaker 1

Building fires in the middle of roads, Polish farmers are once again blocking traffic in nationwide protests.

00:00:07:24 Speaker 1

They've brought their tractors and even a coffin they say symbolizes the death of agriculture.

00:00:14:21

00:00:17:17 Speaker 2

Spring is coming.

00:00:18:24
00:00:19:04 Speaker 2

Figure 24 *Speaker labels with renaming option*

4.3.2. *New Engines or Providers*

Overall, we have investigated and evaluated new providers throughout the project. This was important as new engines came up that covered low-resource languages that did not have any (usable) provider, such as Pashto (one of DW's languages). Once the engine is found to be adequate for DW's use case, it was implemented in plain X by Priberam. The respective editorial department is then informed and asked to evaluate it. This includes, for instance, LESAN engine for Amharic ASR and Whisper3 for Pashto ASR.

Of course, we added specific SELMA engines, such as the AST (automatic speech translation, speech-to-translated text) engine French-English, French ASR, and customized voices for Brazilian Portuguese and Urdu (see section 4.13).

4.3.3 *Subtitling Templates*

Subtitling templates were enabled and enhanced during the final year. This gradually increased from basic settings, to being able to edit and delete the templates and use more settings, including exact positioning and color of font and background.

Users with a higher access level, such as workspace managers, can create, edit and delete the templates. All users can select and apply the templates.

We gathered input from our users, editorial departments and design departments on requirements for approved subtitling settings and design. Certain departments needed customized settings for distribution to other organizations, with, for instance, a different subtitle length (e.g. 32 characters in lieu of the regular 35).

We created, for example, a template to be used when inserts are displayed on the screen. The special template "above inserts" puts the text immediately above the insert, so that the inserts are not covered by the subtitle. The user has the choice of selecting such templates for single segments only or for the entire document. They could decide to use the "above inserts" template for the entire video, to avoid the subtitles shifting position. That also saves time, as individual segments do not need to be adapted. This is obviously also determined by the house style.

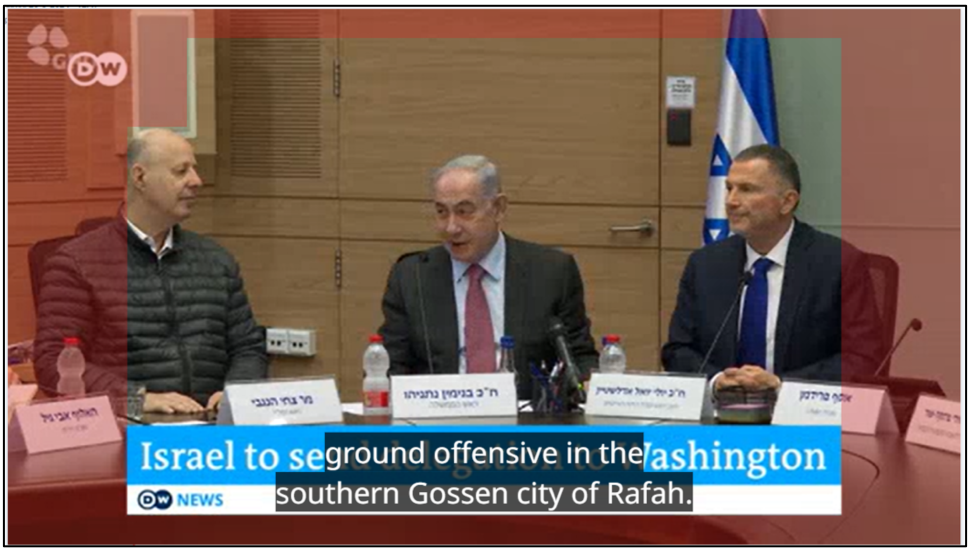
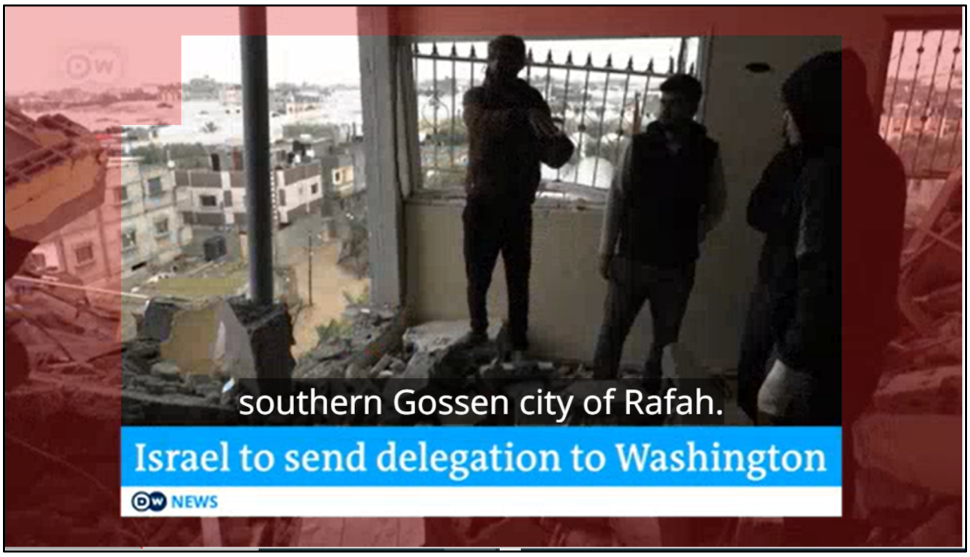


Figure 25 *Subtitle covering the insert*

Below is an example of an instance where the insert is not covered by the subtitle, by using a suitable template.

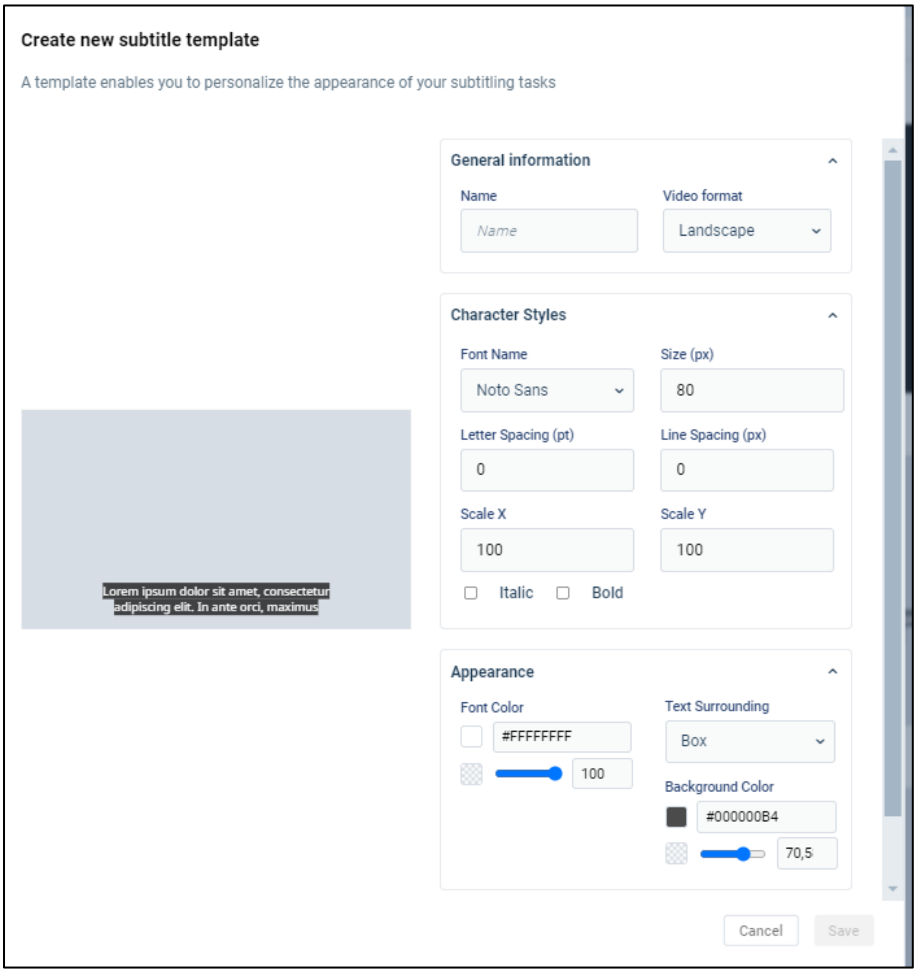


southern Gossen city of Rafah.

Israel to send delegation to Washington

__Figure 26__ One-line subtitle using template “above inserts”

Customized subtitle templates can be created for different types of content, with different styles and positions, targeting different distribution channels. The image below shows the different settings that can be indicated in a subtitle template.



Create new subtitle template

A template enables you to personalize the appearance of your subtitling tasks

Lorem ipsum dolor sit amet, consectetur adipiscing elit. In ante orci, maximus

General information

Name: Name

Video format: Landscape

Character Styles

Font Name: Noto Sans

Size (px): 80

Letter Spacing (pt): 0

Line Spacing (px): 0

Scale X: 100

Scale Y: 100

□ Italic □ Bold

Appearance

Font Color: #FFFFFFFF, 100

Text Surrounding: Box

Background Color: #000000B4, 70,5

__Figure 27__ Subtitle template settings

Below we show a number of templates that were created based on DW requirements, after communications with editorial and design departments. Users can now simply select one of the templates when editing the subtitles in the platform. It is displayed in the plain X platform according to the settings. This has resulted in improved quality of the subtitle output and time savings for the editor. Whether this then in the end has the same format when exported depends on the subtitle export format and the player.

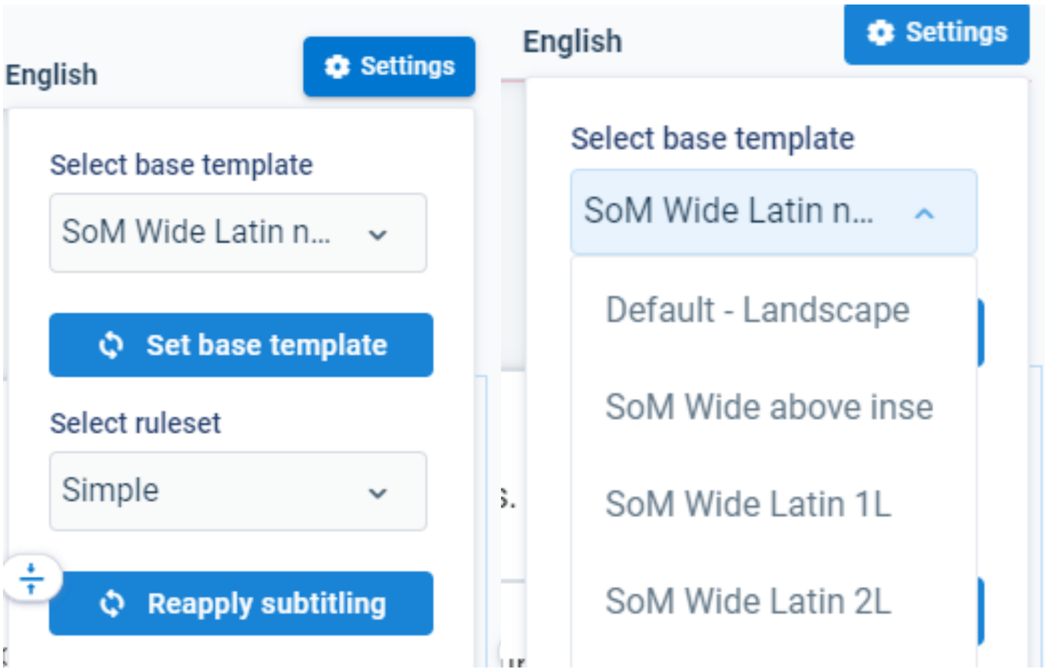


Figure 28 *User selecting a template to be applied*

In practice, the workspace manager creates templates specifying certain parameters, and the editor can then select the appropriate template.

4.3.4 Subtitling Quality

Another enhancement in the use of subtitles is subtitle rules and settings applied to the segmentation.

User feedback and design requirements resulted in a set of rules that should be applied in subtitle segmentation, which was gradually expanded during the project. This included, for instance, parameters as to the length of the subtitle lines, the number of characters, depending on the language and format (horizontal, square and vertical videos), avoiding splitting between first and last name, or title and name (Mr. Brown, Dr. Philips). Title/acronym punctuation should be distinguished from end-of-sentence punctuation, only the latter should trigger a new subtitle line or segment.

This resulted in changes in the code to improve the layout and structure of the subtitles. This is shown to the user as an option between a simple and a grammar ruleset. We have included this in the training sessions for the users and explain the difference.

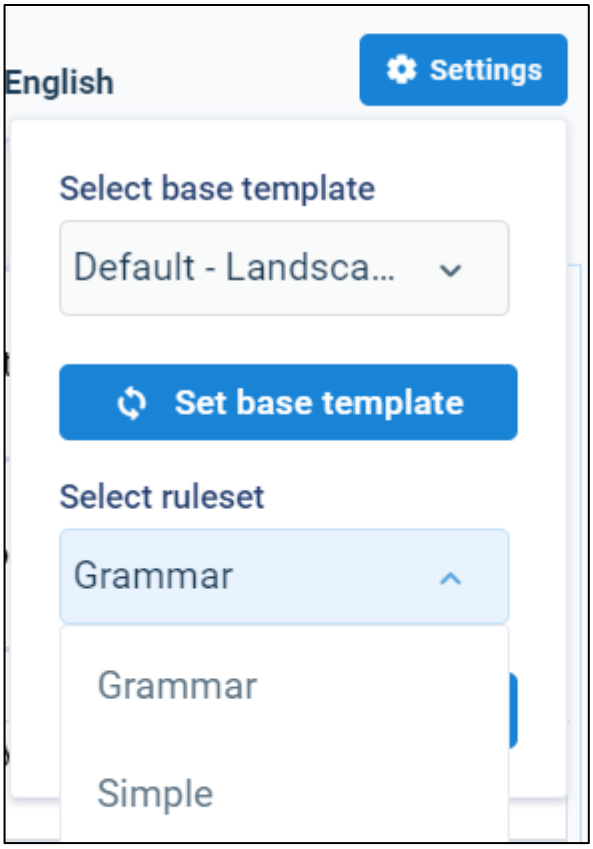
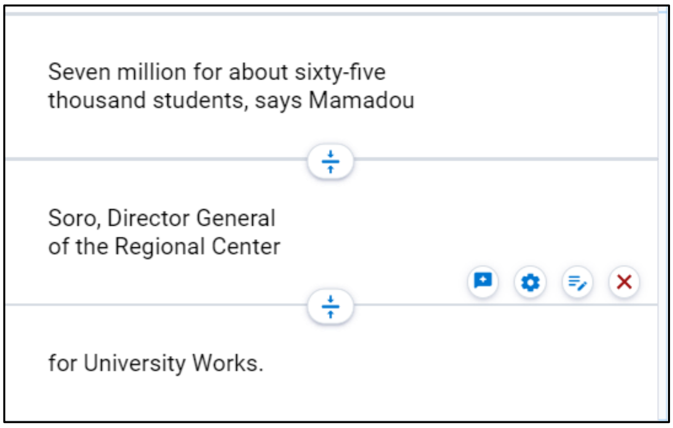


Figure 29 *User choosing between simple or grammar-enhanced subtitling*

We have made a comparison between the results of the two. Below is an example, showing that the grammar option improves the way some named entities are displayed. It is more structured and a combination of several preset rules and a grammar-enhanced algorithm. With the simple method, the name *Mamadou Soro* is split and so is the title *Director General of the Regional Center for University Works*. Using the grammar-enhanced option, the names and titles are displayed in a more consistent way.



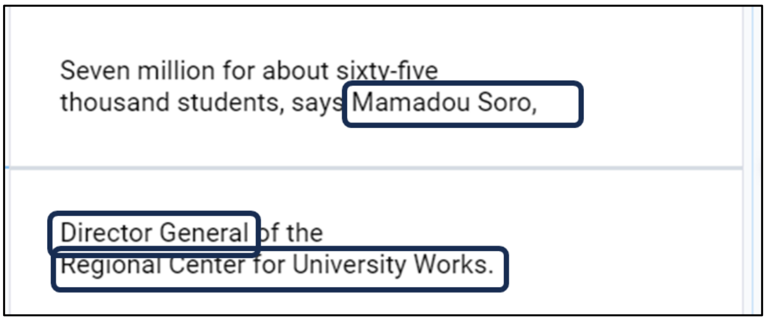
Seven million for about sixty-five
thousand students, says Mamadou

Soro, Director General
of the Regional Center

for University Works.

Figure 30 *Simple subtitle segmentation*

Below is an improved version of the subtitles, using grammar-enhanced segmentation:



Seven million for about sixty-five
thousand students, says Mamadou Soro,

Director General of the
Regional Center for University Works.

Figure 31 *Grammar-enhanced subtitle segmentation*

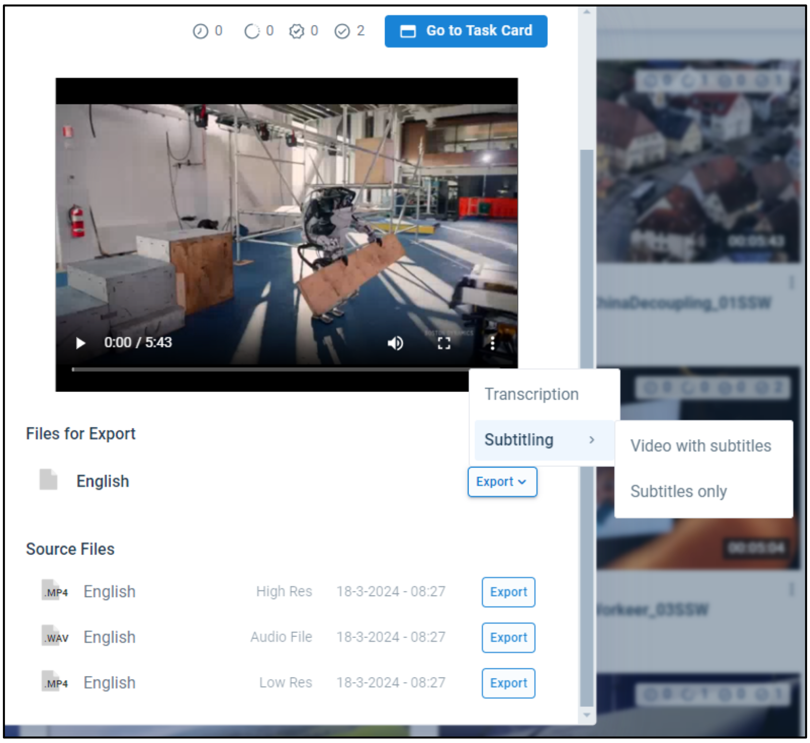
Thus, grammar-enhanced subtitling gives overall better results. These have to be combined, however, with house-style specific rules, such as DW's rule that the subtitle should be displayed as a pyramid, so the second line should be longer than (or equal to) the first line. These two sets of rules (grammar vs house rules) sometimes collide. Yet, it is extremely important to generate the subtitles automatically in the best possible way, adhering to the house style, yet showing consistency in terms of named entities and grammar, and keeping the need for human post-editing to a minimum, preferably down to nill. It should result in a better and more customized subtitling output.

4.4.5 Subtitling Export Formats

Initially, srt was the only subtitling format. This was good enough at the beginning, as srt subtitles were imported into a post-processing and publishing platform and burnt in. However, as more users were active and more subtitles were processed, it became apparent that other publishing channels and formats were needed. Closed captions are preferred, so one video can have subtitles on demand in different languages. Also, srt does not allow any positioning or other information, only text and timecodes. Formats such VTT and EBU-STL include information on positioning (e.g. above the inserts or at the top of the page), color (colored speaker labeling for enhanced accessibility).

Thus, after exchange with the users from various organizations, the following subtitle export formats are now available to the user:

- Subrip (.srt)
- WebVTT (.vtt) - With or without styles
- Plain text (.txt)
- AVID (.avid)
- EBU-STL (stl)
- SRT
- AdvSS (.ass)



Files for Export

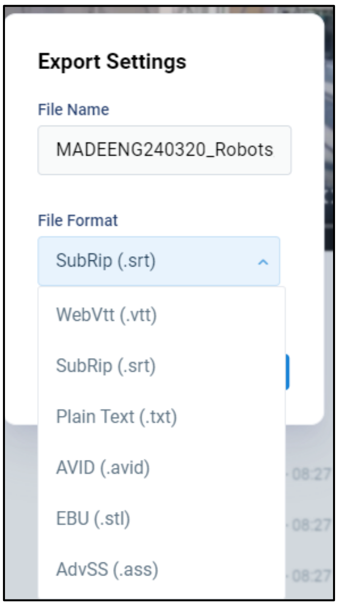
English

Source Files

.MP4	English	High Res	18-3-2024 - 08:27	Export
.WAV	English	Audio File	18-3-2024 - 08:27	Export
.MP4	English	Low Res	18-3-2024 - 08:27	Export

Figure 32 *The user chooses export options*

Different export formats are shown below:



Export Settings

File Name: MADEENG240320_Robots

File Format: SubRip (.srt)

Figure 33 *Several subtitle export formats, resulting from user feedback*

At DW, we currently use different formats simultaneously, srt is used as the standard output format for the content system, EBU-STL is used to forward to external broadcast partners, VTT is used for publication to the dw.com website through the CMS. The default in the platform is VTT with styles, but DW has requested VTT without styles, as styles are incompatible with the current player settings. If we use VTT without styles or srt, positioning information cannot be transferred and is handled by the publishing system. To enable these enhanced settings such as positioning, styles that are compatible with the player need to be determined.

4.4.6 Accessibility

Enhancing access to and through the plain X tool for people with disabilities has been an important aspect from early days on.

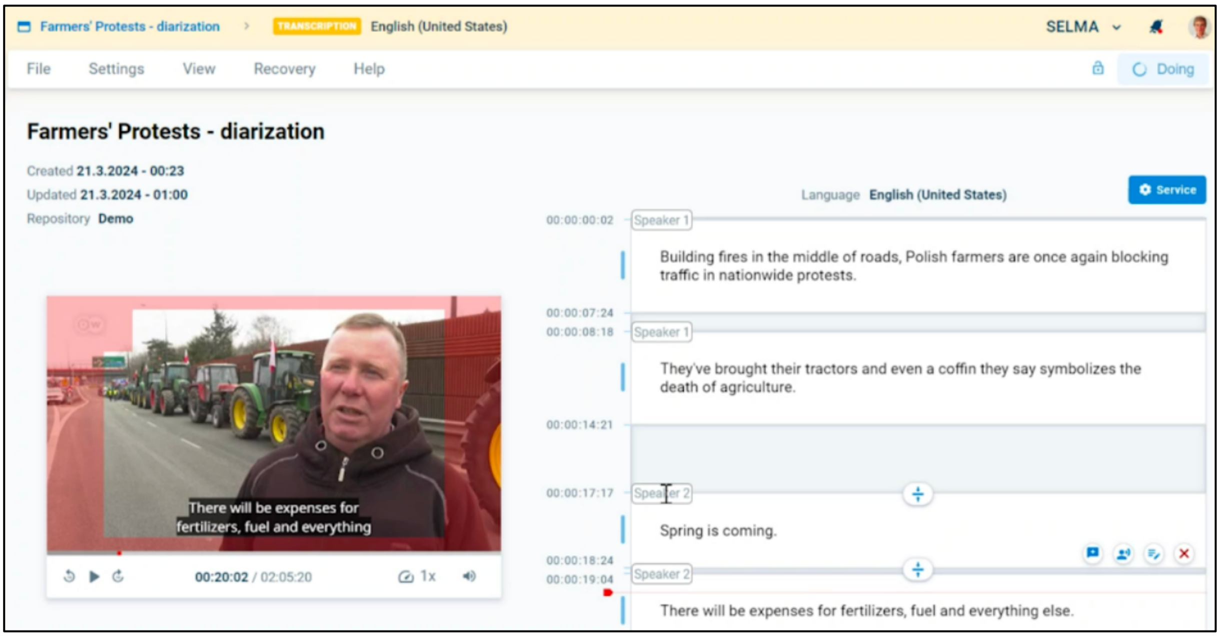
In terms of access to the tool, plain X uses two standards: WCAG and ARIA. WCAG (The Web Content Accessibility Guidelines) is a shared standard for web content accessibility for individuals, organizations, and governments. Currently, accessibility tests show, that plain X is covering these accessibility features:

- Keyboard Accessibility
- Text Alternatives (Alternative Text) for non-text content (images, video and audio)
- Captions for videos with audio
- Audio description or text transcript for videos with sound and pre-recorded videos
- Orientation - the platform adapts to portrait and landscape views
- Focus indication for interactive elements - inputs and modals; focus order
- Semantic HTML applied to headings
- Color Contrast
- Modals follow ARIA patterns
- Dropdowns, selectors, search, and menu bars are WAI-ARIA compliant
- Character Key Shortcuts

Concerning access through the tool, plain x is an Accessibility tool per se, since it creates subtitles in a semi-automated manner for vast amounts of data in many languages. DW has already committed itself to produce up to 100% of subtitles for all their content in all languages by 2025 based on plain X.

Future plain X accessibility scenarios to test, refine and evaluate include the following aspects:

- Automated speaker labeling through diarization (quite advanced status)
- Colored speaker labeling possible in certain export formats (VTT, EBU-STL, for example), using templates
- Automated voicing in other languages (serving cultures and regions which primarily consume media in an oral way as well as vision impaired users speaking another language)



Farmers' Protests - diarization

Created 21.3.2024 - 00:23
Updated 21.3.2024 - 01:00
Repository Demo

Language English (United States)

00:00:00:02 Speaker 1

Building fires in the middle of roads, Polish farmers are once again blocking traffic in nationwide protests.

00:00:07:24
00:00:08:18 Speaker 1

They've brought their tractors and even a coffin they say symbolizes the death of agriculture.

00:00:14:21

00:00:17:17 Speaker 2

Spring is coming.

00:00:18:24
00:00:19:04 Speaker 2

There will be expenses for fertilizers, fuel and everything else.

***Figure 34** plain X with diarization and speaker labeling (Speaker 1, Speaker 2)*

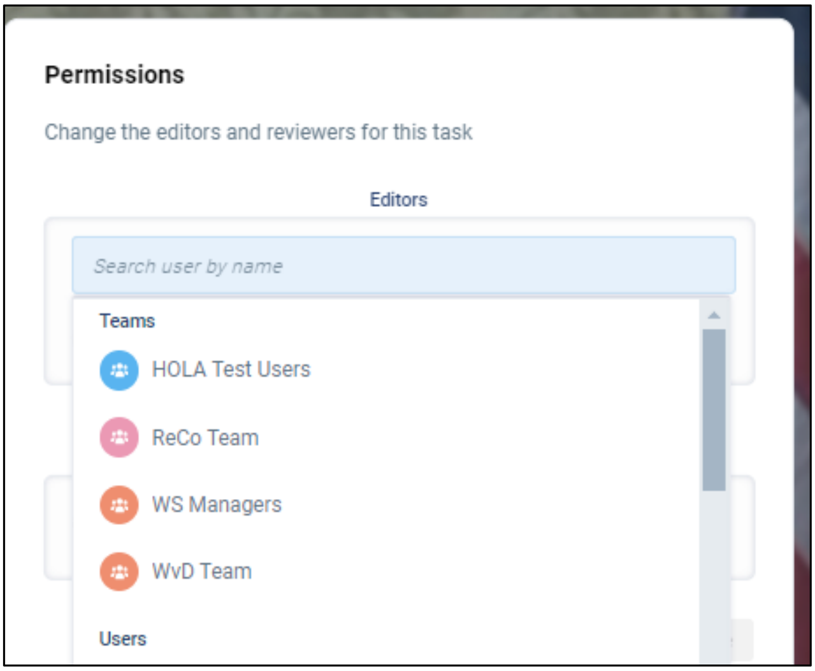
4.4.7. Collaboration

Another feature that was added based on user demand after having used the platform is that the current setup protects the editor's privacy, but is too restrictive in terms of sharing content and tasks. Task sharing was already possible, but each individual had to be specifically added to the task.

The platform was set up to ensure that the editors have full control of their content and the tasks they are working on, but if they become unavailable unexpectedly, the tasks cannot be assumed by a colleague without a workspace manager or team leader having to reassign them to another person.

Thus, after discussions with editorial teams and management, it was decided to change the system so that tasks can now be assigned to individuals and to teams. This means the department has the option to open up the task to their entire team and they appoint someone internally to perform the task. This gives everyone full flexibility, from one individual getting access to an entire team.

We have also started using the guest editor function, whereby tasks can be assigned to remote external editors, for instance native speakers in Asia. This leads to an expansion of the workforce, a reduction in cost, and the possibility to get tasks done faster.



Permissions

Change the editors and reviewers for this task

Figure 35 *Tasks can now be assigned to teams as well as individuals*

4.4.8 User Testing Process Enhancement

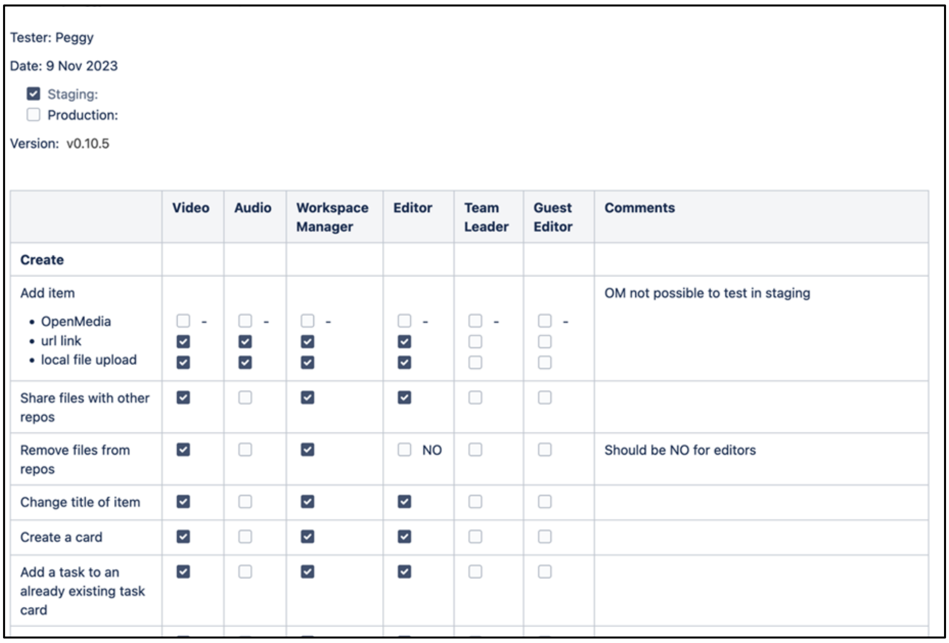
With several active instances of plain X, including a live operational platform connected to DW infrastructure, continuous user testing is essential. We therefore enhanced the testing and assessment process, to make it consistent and reliable.

As users, we routinely test the plain X platform to ensure that it is free of bugs and that new features work as expected. The testing needs to happen every time the platform is updated. This can mean that the system needs to be tested extensively multiple times per month by different users. As such, we need to ensure that the testing is consistent across users, and the output is recorded properly.

Test Sheet

To do so, we have compiled a test sheet that describes the different steps to be tested in the platform. Each user then chooses a role in the platform to test (e.g. editor or team lead) and follows each step as described in the sheet. They then ensure that the feature works as intended or make a note of the problem. Once the test sheet is filled in completely, the form is saved in a cloud system and the bugs are passed to the developers.

The form presents in the form of a table, with features to be tested divided in 5 general groups: the general functionalities of the platform (e.g. adding an item, creating a team, inviting users) and the functionalities of a specific workflow (e.g. creating a transcription, a translation, subtitles and voice-over). The form also differentiates between the different user roles, as this impacts permissions and thus the availability of features. The test sheet is filled in for each new update.



Tester: Peggy

Date: 9 Nov 2023

☑ Staging:
☐ Production:

Version: v0.10.5

	Video	Audio	Workspace Manager	Editor	Team Leader	Guest Editor	Comments
Create							
Add item • OpenMedia • url link • local file upload	 ☐ - ☑ ☑	 ☐ - ☑ ☑	 ☐ - ☑ ☑	 ☐ - ☑ ☑	 ☐ - ☐ ☐	 ☐ - ☐ ☐	OM not possible to test in staging
Share files with other repos	☑	☐	☑	☑	☐	☐	
Remove files from repos	☑	☐	☑	☐ NO	☐	☐	Should be NO for editors
Change title of item	☑	☐	☑	☑	☐	☐	
Create a card	☑	☐	☑	☑	☐	☐	
Add a task to an already existing task card	☑	☐	☑	☑	☐	☐	

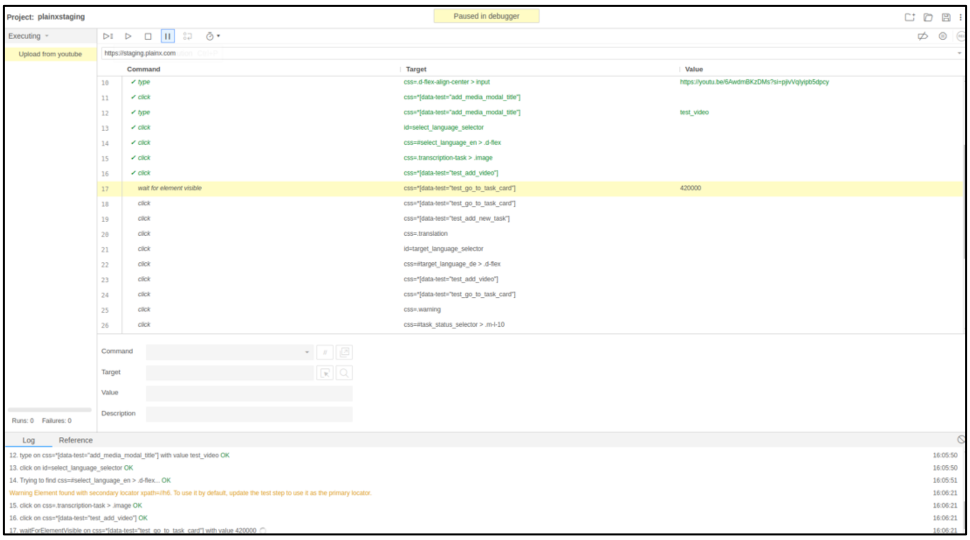
***Figure 36** plain X Test Sheet*

Automated Test Process

A framework comprising functional tests was established with the objective of identifying defects and inconsistencies in the application's behavior. Through the systematic examination of various functionalities and scenarios, functional tests play a crucial role in revealing bugs, usability issues, and edge cases that may otherwise remain undetected. By addressing these issues early in the development lifecycle, we can deliver higher-quality software that aligns with user expectations and requirements.

Moreover, this test suite significantly reduces the time and effort required for manual testing. By automating repetitive test cases, such as video uploads on the platform and the generation of automated transcripts and translations, tests can be executed more efficiently and swiftly. This, in turn, allows for the allocation of freed-up time towards conducting usability testing, thereby enhancing the overall quality of the software product.
We have opted to utilize Selenium as our testing framework due to its open-source nature and cross-browser compatibility. Employing Selenium primarily for browser automation tasks, we've streamlined processes like signing in, uploading videos from both YouTube and local

machines, selecting transcription, translation, and voice-over services, and subsequently downloading transcripts in various formats. To maintain consistency in our tests, we consistently use the same video. Following test execution, a report is generated detailing the number of tests that passed and failed, along with any exceptions thrown. This systematic approach enables us to compare results with previous tests and assess the efficacy of our processes.



	Command	Target	Value
10	✓ type	css=.d-flex-align-center > input	https://youtu.be/6AwdmBKzDMs?si=pjivVqlyipb5dpcy
11	✓ click	css=*[data-test="add_media_modal_title"]	
12	✓ type	css=*[data-test="add_media_modal_title"]	test_video
13	✓ click	id=select_language_selector	
14	✓ click	css=#select_language_en > .d-flex	
15	✓ click	css=.transcription-task > .image	
16	✓ click	css=*[data-test="test_add_video"]	
17	wait for element visible	css=*[data-test="test_go_to_task_card"]	420000
18	click	css=*[data-test="test_go_to_task_card"]	
19	click	css=*[data-test="test_add_new_task"]	
20	click	css=.translation	
21	click	id=target_language_selector	
22	click	css=#target_language_de > .d-flex	
23	click	css=*[data-test="test_add_video"]	
24	click	css=*[data-test="test_go_to_task_card"]	
25	click	css=.warning	
26	click	css=#task_status_selector > .m-l-10	

```
12. type on css=*[data-test="add_media_modal_title"] with value test_video OK      16:05:50
13. click on id=select_language_selector OK                                        16:05:50
14. Trying to find css=#select_language_en > .d-flex... OK                         16:05:51
Warning Element found with secondary locator xpath=//h6. To use it by default, update the test step to use it as the primary locator.   16:06:21
15. click on css=.transcription-task > .image OK                                   16:06:21
16. click on css=*[data-test="test_add_video"] OK                                  16:06:21
17. waitForElementVisible on css=*[data-test="test_go_to_task_card"] with value 420000   16:06:21
```

Figure 37 *Automated plain X Testing*

Feedback process

After each deploy, the platform is tested extensively following the test sheet previously described. All detected bugs or missing features are then compiled and translated into tickets to pass on to the development team. For this, we have used the platform Trello. It allows us to assign tickets, prioritize them, add a detailed description as well as archiving them.

Each ticket was discussed during weekly meetings between DW and Priberam.

On DW’s side, users reported bugs and feature requests through Teams.

4.4.9 Integration

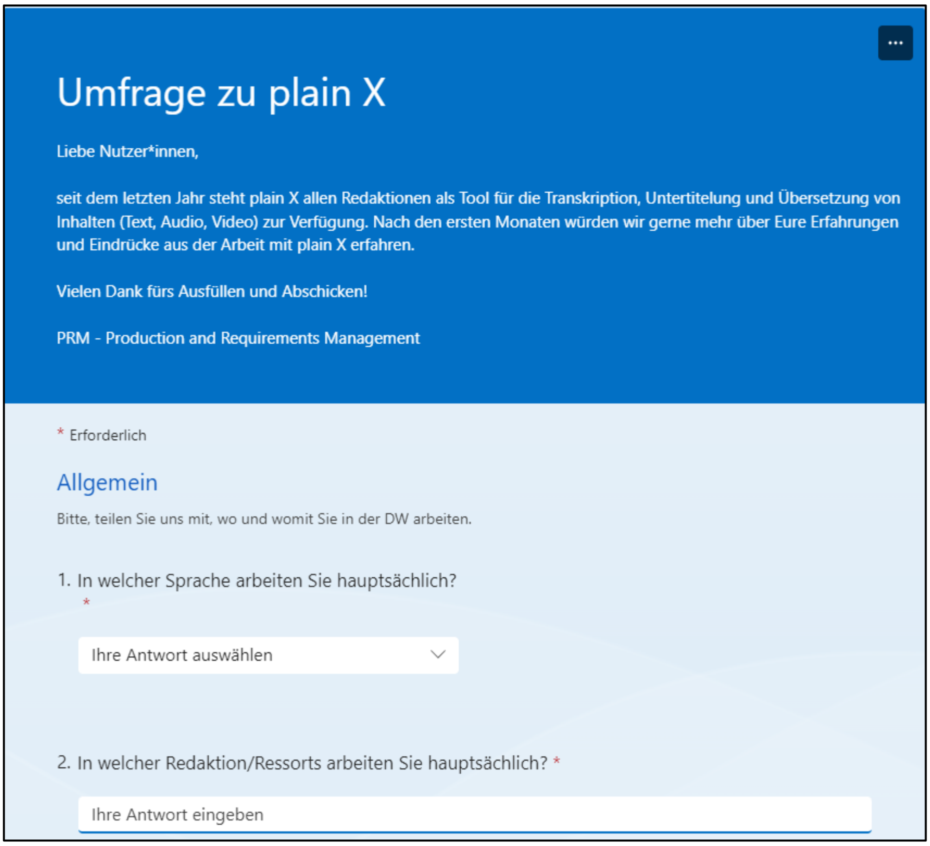
The plain X team and editorial and technical teams in DW have worked together intensively to ensure the NLP technologies can be used efficiently in-house and connect smoothly with other systems, including Hive, Premiere, CMS and OpenMedia.

Through continuous testing, adaptations and retrials, content is now imported and exported:

- We can easily export content from OpenMedia (the editorial content system). With a single button, video and corresponding manuscript are sent to plain X. A status report indicates if the item was successfully exported – also if it was subsequently deleted from plain X.
- When a transcript from an item coming from OpenMedia is processed in plain X and set as "done", the edited transcript is automatically added to the OpenMedia repository in DW
- When a subtitling file is processed in plain X and set "as done", its srt version is automatically sent to the Hive internal system
- When a subtitling file is set as "done", a VTT without styles file can be downloaded and uploaded to the CMS. This will be further automated in the future.
- An srt can be exported and easily imported in Premiere for post-processing, for instance to do video editing or to add inserts. This will also be further automated.

4.4.10 User Satisfaction Questionnaire

In order to get feedback on the use of plain X within Deutsche Welle, we sent out questionnaires to the somewhat 600 currently registered users in the DW instance of the platform, and so far got back 77 responses.



Umfrage zu plain X

Liebe Nutzer*innen,

seit dem letzten Jahr steht plain X allen Redaktionen als Tool für die Transkription, Untertitelung und Übersetzung von Inhalten (Text, Audio, Video) zur Verfügung. Nach den ersten Monaten würden wir gerne mehr über Eure Erfahrungen und Eindrücke aus der Arbeit mit plain X erfahren.

Vielen Dank fürs Ausfüllen und Abschicken!

PRM - Production and Requirements Management

* Erforderlich

Allgemein

Bitte, teilen Sie uns mit, wo und womit Sie in der DW arbeiten.

1. In welcher Sprache arbeiten Sie hauptsächlich? *

Ihre Antwort auswählen

2. In welcher Redaktion/Ressorts arbeiten Sie hauptsächlich? *

Ihre Antwort eingeben

***Figure 38** plain X User Satisfaction Questionnaire*

We include here the most relevant questions as to the user satisfaction of plain X.

One question addressed the primary usage of plain X. It asks if they use plain X primarily for transcription (Transkription), translation (Übersetzung) or subtitling (Untertitelung), As we can see, most of the editors use it for transcription, then subtitling and finally translation.

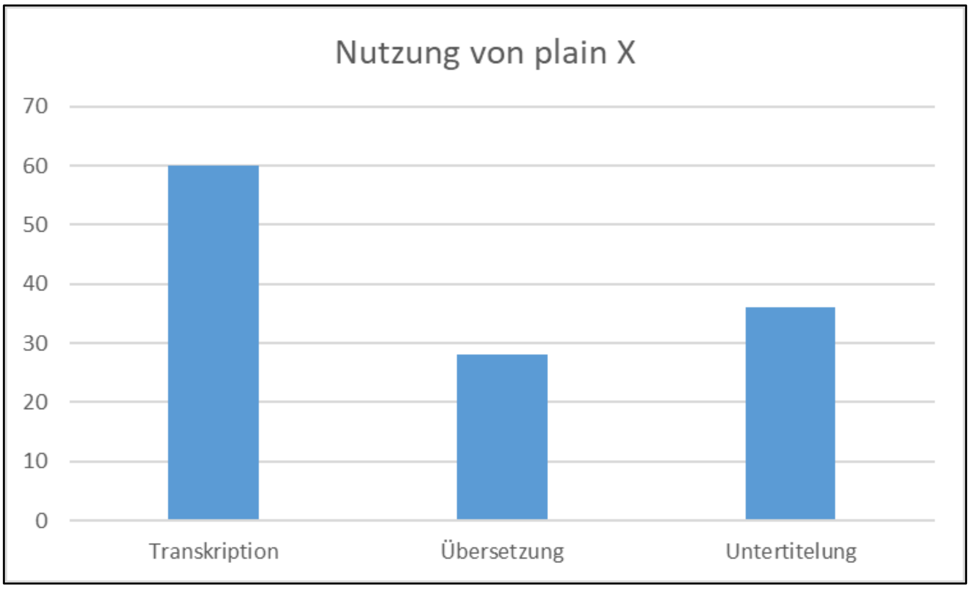


Figure 39 *Primary usage of plain X*

The next three graphs give us insights into the satisfaction using transcription, translation and subtitling in plain X. It uses a four-level Likert scale:

- 1 auf jeden Fall = absolutely
- 2 auf keinen Fall = definitely not
- 3 eher nicht = rather no
- 4 eher schon = rather yes

The following question inquired about their satisfaction with plain X for transcription. Most users were quite happy with that function.

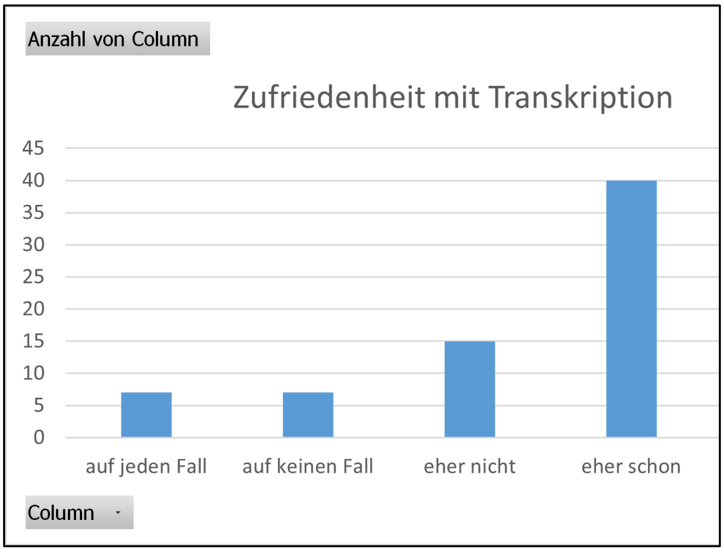


***Figure 40** plain X Satisfaction with Transcription*

The next question asked about their satisfaction as to translation. Here the responses are more mixed, understandably, as translation is much more subjective and needs post-editing in most cases, and definitely in low-resourced languages. It does mean that we need to look for better engines in low-resourced languages, or improve the output with supplementary modules. And ensure that we guide the users in which engines get the best results, based on our benchmarking. Guiding the users is through setting default engines for the different processes.

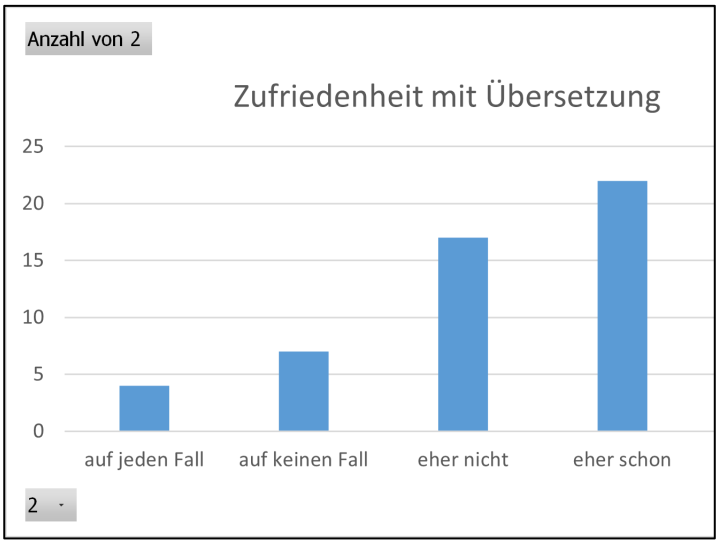


***Figure 41** plain X Satisfaction with Translation*

Finally, the graph below asked about user satisfaction with the subtitling. The vast majority is quite satisfied with the subtitling.

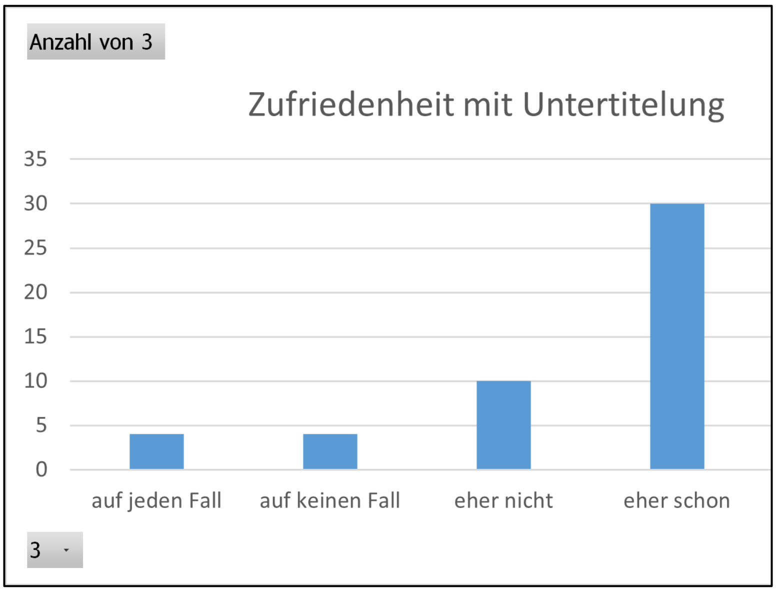


***Figure 42** plain X Satisfaction with Subtitling*

Next, we provide some additional feedback that came out of the questionnaire.

Comments and suggested improvements on **transcription**:

Issues:

- It is very important for the transcription to be able to distinguish the speakers. The tool is often unable to do this. It is very time-consuming to make this assignment manually, and there are often problems with the automatically generated timestamps.
- Sometimes plainX recognizes some words incorrectly or sets the period incorrectly.
- The spacing. Some inserts only last 1 second and I always have to extend the duration manually.
- Sometimes the translation/transcription from Arabic is not correct. I know that Arabic is a very complicated language, but I trust that plain X could be developed further
- The texts are translated incompletely, whole paragraphs are missing. It is much too slow. I no longer use it after 2 agonizing attempts, as the transcription on Premiere is extremely much faster.

- Many words are not transcribed correctly. Questions and answers are mixed. Without watching the videos, the transcription is completely incomprehensible.
- Punctuation
- Precision, e.g. when recognizing languages other than the set language (e.g. beginning and end of foreign-language O-tones)
- Accuracy
- Only one transcription language can be selected at a time -> problematic with multilingual film material
- Transcripts always have to be laboriously edited (words not correctly recognized, assignment of the various speakers not precise enough, etc.)
- Speed
- It is an exception, but one interlocutor speaks very slowly and the tool interprets this as a sentence change or point. Perhaps the tool can be recognized algorithmically at some point.

Suggestions:

- Add export into a Word file
- Recognize input names (countries, cities, organizations), distinguish voices and match them correctly
- Manual timecode adjustment, speech2text based on the manuscript and AI (without manuscript) in one process. E.g. through IN and OUT

Comments and suggested improvements on **translation**:

Issues:

- Some languages are translated literally, which distorts the meaning. Colloquial language is sometimes translated in an extremely incomprehensible way.
- The quality really depends on the language.
- Overall, we need a better translation quality.

Comments and suggested improvements on **subtitling**:

Issues:

- Timecode setting. It's already relatively good for English - but not good enough to work without a reviewer. There is still more catching up to do in the other languages.
- The playback feedback should be more precise to make it easier to adjust the timecodes.
- Position of the CC for vertical videos need to be improved
- Most importantly, I would like to see an improvement in the manual adjustment of subtitles. A shortcut for skipping back and forth in individual frames would be very practical. It would also be very nice if the player display were more synchronized with the subtitles. Until now, you have to guess where an original soundtrack begins or ends and can't rely on the displays.
- As I never deal with just one video, but usually a whole series of videos, I think that the manual adjustments that are still necessary should be reduced.
- Handling is difficult. Even with subtitles - the control symbols are very small and difficult to respond to.

Suggestions:

- Position of the captions on vertical videos should be flexible to change
- Add a shortcut to move through the video frame by frame while editing the subtitles. And please also add an audio curve so you can better see where the sound starts and stops without having to constantly replay the video. That would make editing immensely easier and I really miss that about the WinCaps program we used for subtitling before plain X was introduced.
- Also a batch processing would be interesting.

As we can see from the satisfaction graphs, users are overall quite happy with the platform, in particular for transcription and subtitling. Translation is harder to meet expectations or the quality they aim at, as translation is very subjective and hard to automate. It will never translate it as the translator would do it. This is very different from transcriptoin. Therefore, expectation management continues to be a communication we have to work on.

As for the suggestions and comments, this feedback will definitely be taken into account for further development. Some of these issues have already been addressed. Some are a user or a network issue, but we will certainly take some of these points with us to further enhance the system - and look for solutions.

4.4.11 User Guide

Guiding a large number of users, especially with varying use cases and levels of technical expertise, is a fairly challenging, but extremely important task if we want the users to make the most out of such a powerful tool and to use it efficiently.

The DW plain X team has set up an extensive user guide and puts a lot of effort in maintaining it, with updates after every new deploy or in case user inquiries show that a certain feature or action is not clear to the users.

The user guide was initially set up for the DW user group, made available internally through Confluence, and explains all components and features in the tool, as well as Deutsche Welle practices and integrations with DW infrastructure and systems. Advanced features such as collaboration and review processes, preparing and uploading manuscripts, and exporting the appropriate formats for specific publication channels are explained.

User feedback showed that users are sometimes uncertain about the best approach towards certain aspects of the key processes (transcription, translation, voice-over and subtitling). Therefore, a specific section on workflows was added, describing the same processes and tasks in the platform but from a different, less technical point of view.

The user guide was also exported from the internal distribution channel as a Word file and adapted to be used by users external to DW. It was provided, for instance, to external users of the DW Academy, an international training institute for media professionals, and a version was prepared for use by (potential) external clients. A copy of the external user guide is attached as an annex to deliverable D2.7 (Final progress report on continuous massive stream learning).

Conclusion: A lot of work has gone into the development of plain X during the entire project duration and definitely also the final year. We have succeeded to build a strong platform that can do all four processes (transcription, translation, subtitling and voice-over) with relative efficiency.

The work will continue, making the platform more user-friendly, and continue to improve the engines or build modules on top of them, and add new providers and engines to get a better quality of transcription and translation. We will continue to keep gathering feedback from the users to handle change and expectation management and to enhance the platform. Automating the processes takes time and requires close communication with the users.

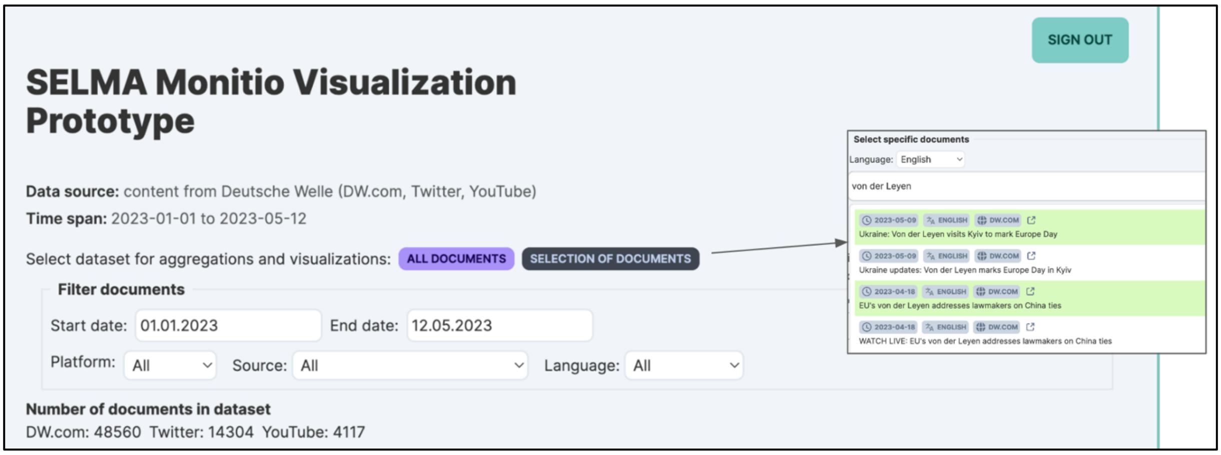
4.4 Diversity Use Case Application

We established and described the Diversity Use Case Application in **Year 1.** This is an application of the news monitoring use case and assesses the ability of the Monitio platform to analyze an arbitrary group of articles with respect to the diversity of their content. We agreed upon the metadata that needed to be added and the source and limitations, i.e., use of only Wikidata for entities and their added metadata, and a suggestion of fields to be added in the UI.

In **Year 2**, this materialized in the form of adding a diversity category in the platform, for named entities, based on metadata from Wikidata. The following fields are added where available: sex and gender, country of citizenship, ethnic group, sexual orientation, medical condition, religion, educated at, date of birth. These fields can be used for filtering and viewing available information. The intended output is statistics on gender (and other minority groups) balance.

The evaluation focused on obtaining accurate and useful statistics. Detailed searches revealed the level of information that can be obtained on these categories and triggered a discussion of what is ethical and permitted in this respect and what the risks are in case of unintended use. The “diversity” fields derived from Wikipedia were reduced to 3 specific ones: binary gender (only female and male), age and birthplace / geographical origin.

In **Year 3,** the Diversity Balance Indicator was further refined, tested and discussed on several occasions. An option was developed and integrated to also analyze a single use item.



SELMA Monitio Visualization Prototype

Data source: content from Deutsche Welle (DW.com, Twitter, YouTube)
Time span: 2023-01-01 to 2023-05-12

Select dataset for aggregations and visualizations: ALL DOCUMENTS SELECTION OF DOCUMENTS

Filter documents

Start date: 01.01.2023 End date: 12.05.2023

Platform: All Source: All Language: All

Number of documents in dataset
DW.com: 48560 Twitter: 14304 YouTube: 4117

Select specific documents

Language: English

von der Leyen

2023-05-09 ENGLISH DW.COM
Ukraine: Von der Leyen visits Kyiv to mark Europe Day

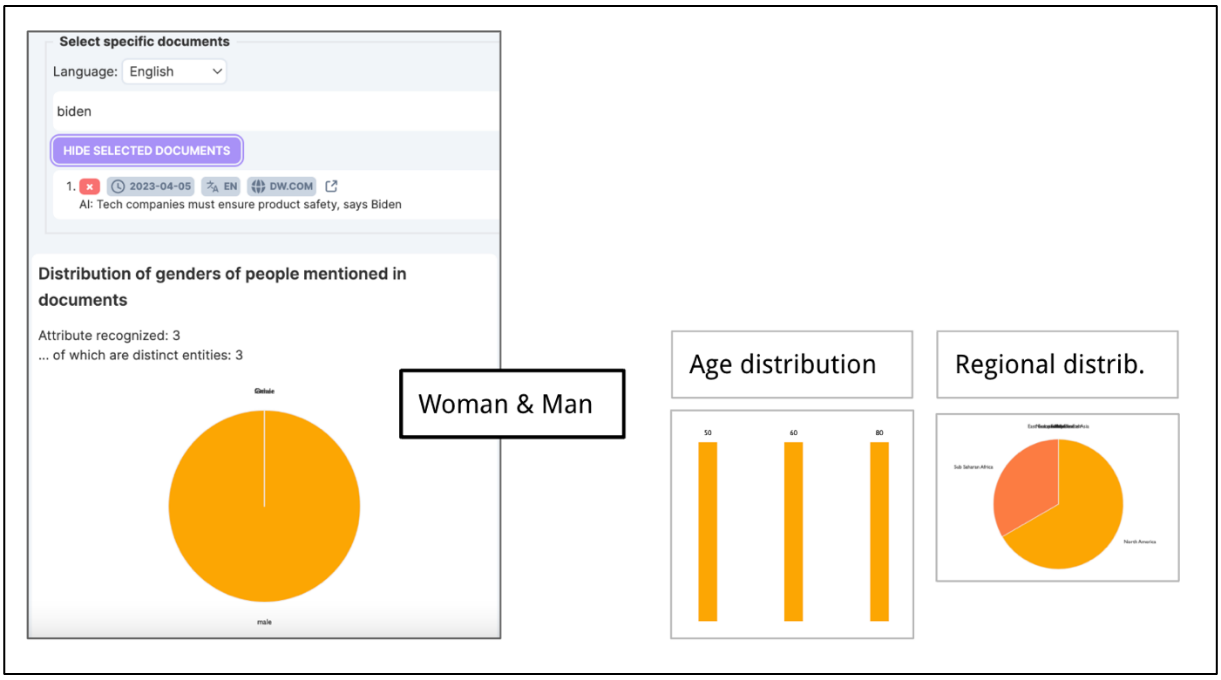
2023-05-09 ENGLISH DW.COM
Ukraine updates: Von der Leyen marks Europe Day in Kyiv

2023-04-18 ENGLISH DW.COM
EU's von der Leyen addresses lawmakers on China ties

2023-04-18 ENGLISH DW.COM
WATCH LIVE: EU's von der Leyen addresses lawmakers on China ties

***Figure 43** SELMA Diversity Balance Prototype with feature to select (and deselect) individual items*

After several negotiation rounds with people from the DW Program department, an editorial department was found to do a longer test. The department with around 15 people tested for a time period of 2 months and used the tool on a daily basis in the context of their editorial conference. The feedback was very positive. It helped them to become aware of the kind of people they usually select. After considering the data (mostly male, mostly older people in their 50ies and above) it also helped them to look for a "new" / more diverse set of protagonists: featuring more female, more diverse (e.g. people with disabilities), also younger voices for their reporting; especially for those news items which were not more or less set by news agencies, but for news reports which they produced themselves.



Select specific documents

Language: English

biden

HIDE SELECTED DOCUMENTS

1. 2023-04-05 EN DW.COM
AI: Tech companies must ensure product safety, says Biden

Distribution of genders of people mentioned in documents

Attribute recognized: 3
... of which are distinct entities: 3

***Figure 44** Showing results of the Diversity Balance Prototype for a single news item*

Conclusion: The overall feedback on the prototype was very positive: the handling of it was easy to learn and the insights gained were considered very valuable. They said that they used it as a "sensitivity" tool, making them aware of how diverse and gender-balanced the people were which they had in their program. Over the time span of 2 months, it contributed to selecting a more diverse set of people. Although the Diversity Balance Indicator can only provide "indications" (only certain attributes are considered including binary gender, age and regional distribution by birth and only people also listed on Wikipedia are part of the analysis; plus: the coverage of Wikipedia differs from region to region - India is not well represented, for example), it helped to get an impression and an idea of “who is in the news”. Obviously, it would be helpful to get a full analysis of all people in the news, not only the ones listed on Wikipedia and to extend the tool with a feature to show development over time.

4.5 Podcast Use Case Application

The Podcast Creator Use Case Application is based on a workflow observed initially in DW's Brazilian language department. The goal of the use case is to increase the workflow's efficiency by supporting the journalist in the production of daily audio news bulletins through SELMA.

This application uses the functionalities of two subordinate applications, viz. the DW Speaker App and the DW Summarizer App described below, to create audio news bulletins. It is a macOS App which was evaluated on MacBook computers running macOS Sonoma.

In **Year 1,** we established the concept for this use case and initiated contact with editorial users that could be involved and set requirements.

The production of a single news bulletin can be subdivided into the following steps. The table shows the duration that is required for each step during the classic, manual process.

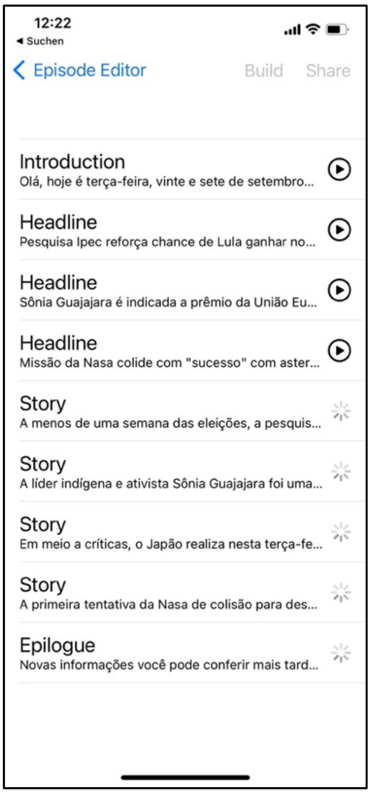
Step	What	Approx. duration
1	Research 5 stories	30 min
2	Write 5 stories	60 min
3	Check stories by colleague	25 min
4	Recording, editing, upload into the system	70 min
5	Add metadata in CMS, create YouTube video, publish on YouTube	45 min
6	Create bi.ly links and publish on Social Media (Twitter & Facebook)	15 min
	Sum	**245 min**

Table 3 *Traditional Process for Podcasting Use Case*

The work in **Year 2** consisted in creating and trying out an automated process to speed up and facilitate some of the steps above. It uses speech technology in the form of DW customized voices developed by LIA, made available through integration with the SELMA OSS, and a newly created Podcast iOS app. This work involved technology partners, such as IMCS and LIA, as well as DW project managers and editors from the Brazilian department, native speakers used to doing the work in the traditional way. These news stories are produced and published twice a day. The SELMA-enhanced module includes a template that streamlines the production process for editorial users and automatically inserts basic, recurring components such as the introduction and music in between the stories.

For this use case, we also assessed the quality of the Brazilian customized voices. More details on this are available in D5.2 section 4.9.

The application already automates the task of creating speech and mixing it with music and background tracks. The following screenshot shows the app while synthesizing speech for the various sections of a Brazilian news podcast by accessing the API provided through UC0.



Introduction
Olá, hoje é terça-feira, vinte e sete de setembro...

Headline
Pesquisa Ipec reforça chance de Lula ganhar no...

Headline
Sônia Guajajara é indicada a prêmio da União Eu...

Headline
Missão da Nasa colide com "sucesso" com aster...

Story
A menos de uma semana das eleições, a pesquis...

Story
A líder indígena e ativista Sônia Guajajara foi uma...

Story
Em meio a críticas, o Japão realiza nesta terça-fe...

Story
A primeira tentativa da Nasa de colisão para des...

Epilogue
Novas informações você pode conferir mais tard...

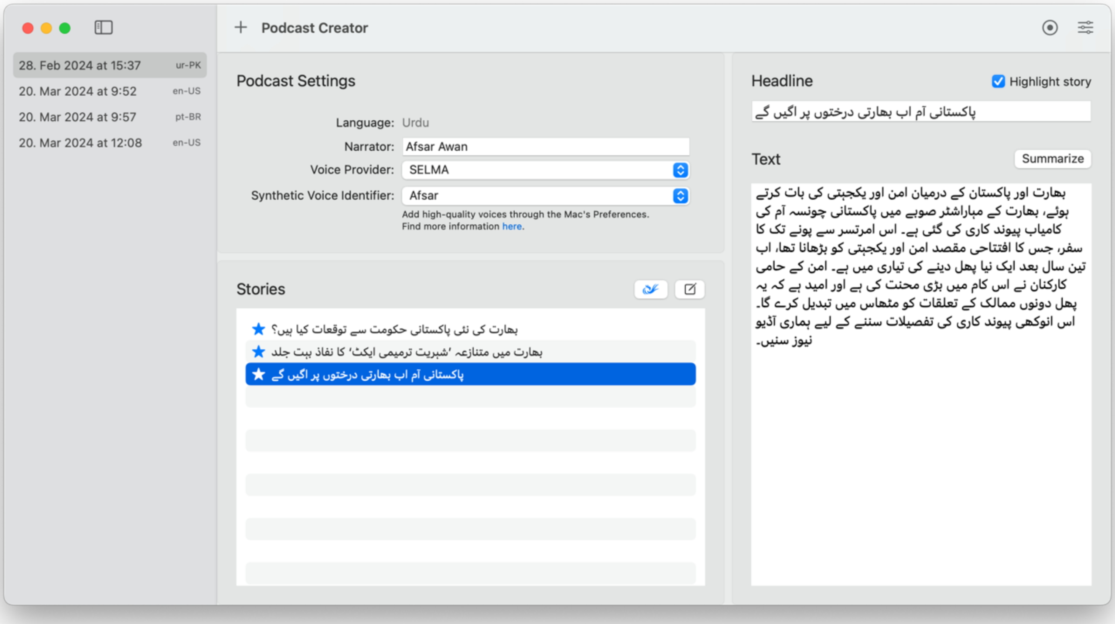
***Figure 45** SELMA Podcast Use Case Automation Template*

The semi-automated process and template was also demonstrated at the SELMA user day workshop in October 2022. Two editorial DW departments, in particular the Hindi and the Urdu departments tried it out, compared it with their current process, and expressed interest in participating in future user trials.

In **Year 3,** further improvements were made to the Podcast Creator app, based on previous evaluations. In particular, the original iOS app was converted into a macOS app, thus making it more accessible to DW's editorial departments. We involved another editorial team, namely the Urdu team, to further try out the automation template for real production, to assess it in terms of gain in effort and productivity.

The following adjustments were discussed, applied and tested in the final stages:

- Integration with the Monitio monitoring platform, resulting in an automation of pre-selecting suitable stories. This was a major change, as it automates the suggestion of storylines for the audio news bulletin to be produced. These clustered stories are generated with SELMA AI modules. We can set the number of storylines, the source language and the time period to be covered for the AI-based clustering. The user can choose the preferred storyline format.
- Enhancements of the summarization itself, so that the journalist has a good starting point for adapting a news story to its audio version
- Enhancements in the audio news bulletin compilation structure.



28. Feb 2024 at 15:37 ur-PK
20. Mar 2024 at 9:52 en-US
20. Mar 2024 at 9:57 pt-BR
20. Mar 2024 at 12:08 en-US

Podcast Settings

Language: Urdu
Narrator: Afsar Awan
Voice Provider: SELMA
Synthetic Voice Identifier: Afsar
Add high-quality voices through the Mac's Preferences.
Find more information here.

Stories

★ بھارت کی نئی پاکستانی حکومت سے توقعات کیا ہیں؟
★ بھارت میں متنازعہ 'شہریت ترمیمی ایکٹ' کا نفاذ بہت جلد
★ پاکستانی آم اب بھارتی درختوں پر اگیں گے

Headline ☑ Highlight story

پاکستانی آم اب بھارتی درختوں پر اگیں گے

Text

بھارت اور پاکستان کے درمیان امن اور یکجہتی کی بات کرتے
ہوئے، بھارت کے مہاراشٹر صوبے میں پاکستانی چونسہ آم کی
کامیاب پیوند کاری کی گئی ہے۔ اس امرتسر سے پونے تک کا
سفر، جس کا افتتاحی مقصد امن اور یکجہتی کو بڑھانا تھا، اب
تین سال بعد ایک نیا پھل دینے کی تیاری میں ہے۔ امن کے حامی
کارکنان نے اس کام میں بڑی محنت کی ہے اور امید ہے کہ یہ
پھل دونوں ممالک کے تعلقات کو مٹھاس میں تبدیل کرے گا۔
اس انوکھی پیوند کاری کی تفصیلات سننے کے لیے ہماری آڈیو
نیوز سنیں۔

***Figure 46** Screenshot of the Podcast Creator*

Finally, the Urdu editorial team assessed the final workflow and output. They looked at it in terms of the main benefits: using the app with a template, using synthetic voices instead of speakers, the entire automated workflow compared to the manual one, having the editor or the monitoring platform select the items, having the editor write the stories or using and editing the SELMA summarization component.

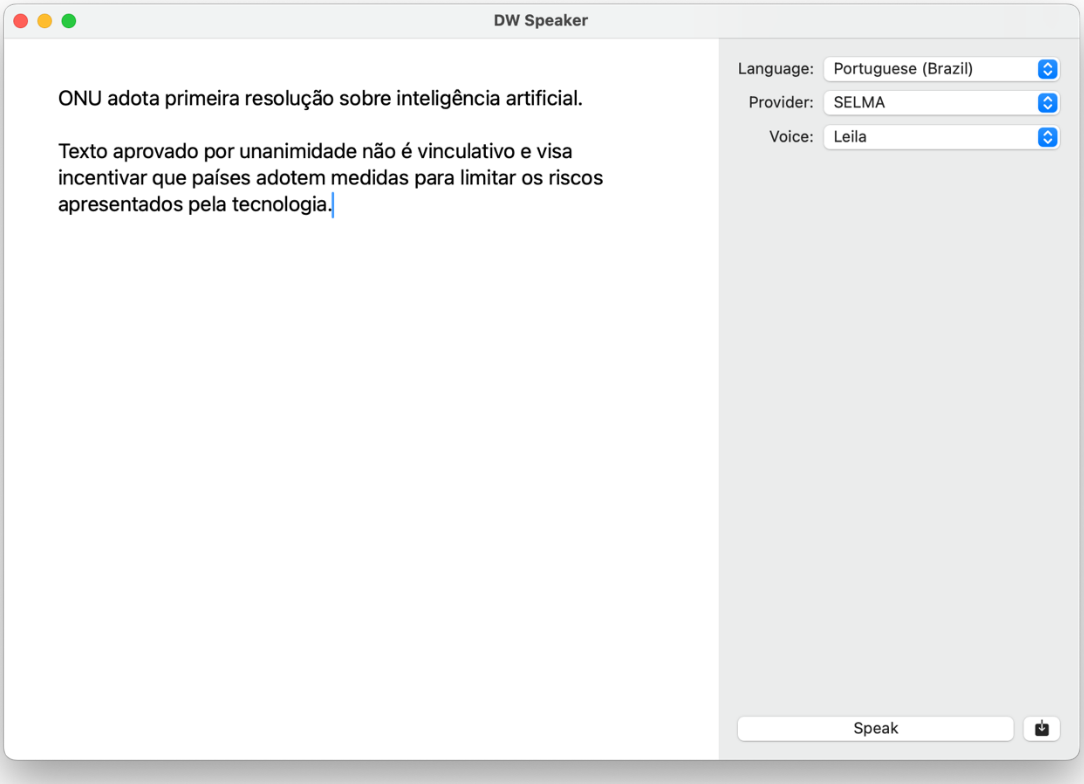
The **conclusion** of this team was that they saw the potential to save a significant amount of time during the production of audio news in Urdu, notably through the summarization, speech synthesis and audio mixing features. They suggested the following further improvements: the ability to display right-to-left text and the possibility to insert audio pauses between headlines and bulletins, thus supporting the separation of the podcasts' segments. The SELMA customized voices for Urdu also need improvements before they can be used in production. Other voices are being considered, possibly in conjunction with voice conversion techniques.

A Deutsche Welle internal pilot project is expected to be launched in the spring of 2024, which will further look at implementing and integrating the podcast creator app in its production systems. It acknowledges the benefits of the tool and establishes what is needed to make the workflows more efficient through the use of the Podcast Creator. Also in focus: what is required for a full integration and what further improvements could be done.

4.6 DW Speaker Application

This macOS application called DW Speaker synthesizes speech from text. A large variety of languages is available from different engines. The app currently includes voices provided through Apple's operating systems, from Google and Azure through the plain X API, as well as SELMA customized voices for Brazilian and Urdu. This macOS App was evaluated on macOS Sonoma.

Through the app UI, text can be copied into the text field and, after selecting the preferred voice and activating the voice button, synthetic speech is rendered to generate an audio file with the spoken text.



ONU adota primeira resolução sobre inteligência artificial.

Texto aprovado por unanimidade não é vinculativo e visa incentivar que países adotem medidas para limitar os riscos apresentados pela tecnologia.

Language: Portuguese (Brazil)

Provider: SELMA

Voice: Leila

Figure 47 *Screenshot of the DW Speaker*

This was the main application through which the DW participating native speakers of Urdu and Brazilian, who also lent their voices for this part of the project, could listen to the outcome of the voices being developed (cloned) and give feedback and suggestions for improvements. After many iterations, we arrived at the current SELMA selection of voices.

The feedback focused on aspects such as tone of voice, volume, interruptions in speech (due to noise, and in particular music, present in the training material that was provided). New versions were provided by the developing partner LIA and assessed repeatedly until reaching the current quality. The conclusion was that such customization/cloning of voices is feasible, but clean training material is highly essential to arrive at good results. The volume of training material needed per speaker is at least 10 hours, with voices sounding increasingly acceptable when more than 20 hours of training material was used.

In **conclusion**, the DW Speaker Application is an easy-to-use tool to use customized synthetic voices to automatically generate an audio file with spoken text. In this project, it helped to have a simple UI focussed on the single task of generating speech to get fast feedback from our users.

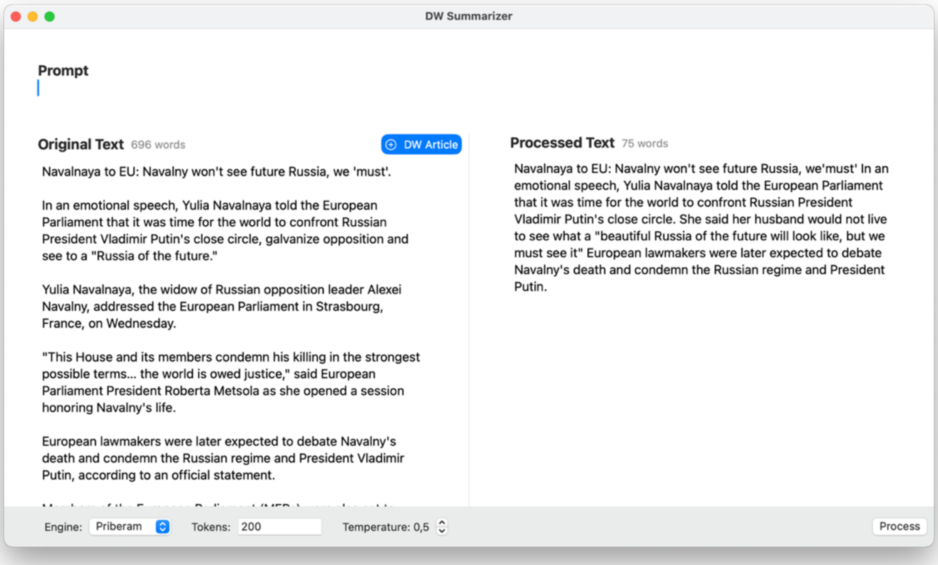
4.7 DW Summarizer Application

The next application is the DW Summarizer, a macOS App that summarizes text. Again, this App was created to run on macOS Sonoma.

The UI is very transparent and smoothly guides the user. We select an engine to be used, drag a content item (for instance an article from the DW website) into the text field. The headline and body of the text appear in the text field. We currently have the choice between ChatGPT Summarization or Priberam's Summarizer, developed in the SELMA project. The first one generates a radio report-style summary and Priberam's tool creates a mono-lingual summary fitting the number of configured tokens.

The assessment consisted primarily of testing the workflow and suggesting enhancements, for instance in terms of structure. We also compared the output of the two service providers currently in the platform (OpenAI ChatGPT and Priberam).

ChatGPT requires a prompt and produces a radio report style summary with three sentences. Priberam's Summarizer does not require a prompt, but works based on the number of tokens in the item.



Navalnaya to EU: Navalny won't see future Russia, we 'must'.

In an emotional speech, Yulia Navalnaya told the European Parliament that it was time for the world to confront Russian President Vladimir Putin's close circle, galvanize opposition and see to a "Russia of the future."

Yulia Navalnaya, the widow of Russian opposition leader Alexei Navalny, addressed the European Parliament in Strasbourg, France, on Wednesday.

"This House and its members condemn his killing in the strongest possible terms... the world is owed justice," said European Parliament President Roberta Metsola as she opened a session honoring Navalny's life.

European lawmakers were later expected to debate Navalny's death and condemn the Russian regime and President Vladimir Putin, according to an official statement.

Navalnaya to EU: Navalny won't see future Russia, we'must' In an emotional speech, Yulia Navalnaya told the European Parliament that it was time for the world to confront Russian President Vladimir Putin's close circle. She said her husband would not live to see what a "beautiful Russia of the future will look like, but we must see it" European lawmakers were later expected to debate Navalny's death and condemn the Russian regime and President Putin.

Engine: Priberam Tokens: 200 Temperature: 0,5

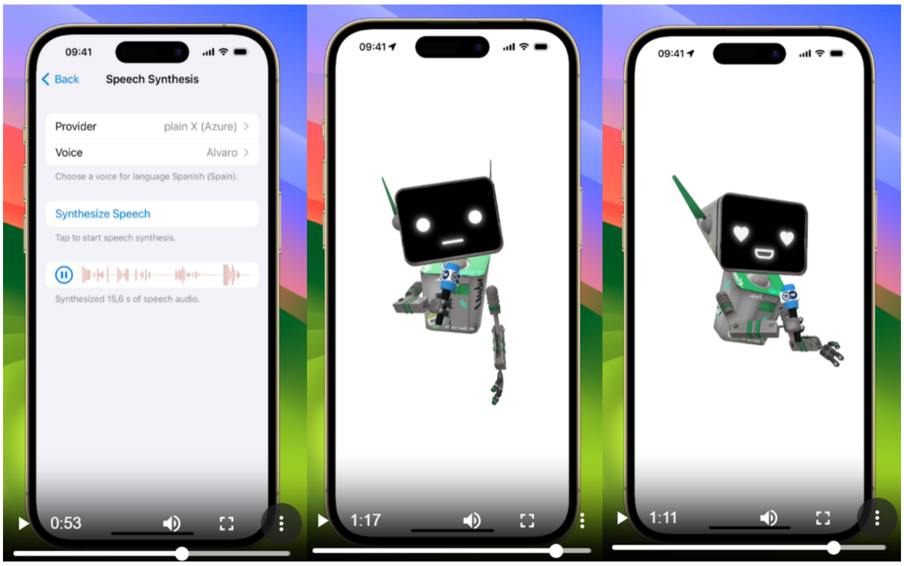
Figure 48 Screenshot of the DW Summarizer

Our user test group's conclusion was that the system works smoothly and is fast and easy to use. The preferred provider is ChatGPT, as it allows journalists to shape the style of the created summary through prompt engineering. Of course, in cases where cost is a concern, as well, the necessity to summarize a large number of documents in privacy-preserving context, Priberam's engine beats the other ones, given that we are able to run the model on our own servers.

In **conclusion**, the DW Summarizer app turned out to be an easy-to-understand vehicle to demonstrate the pros and cons of commercial vs self-developed summarization models.

4.8 DW Avatar Application

A fourth DW app called DW Avatar was created in collaboration with the DW Lab Project "Avatario". A customized DW avatar representing a virtual DW speaker reads news bulletins or other content, using a customized synthetic voice, and moving its head, eyes, mouth and hands in line with the spoken content.



Speech Synthesis

Provider plain X (Azure)

Voice Alvaro

Choose a voice for language Spanish (Spain).

Synthesize Speech

Tap to start speech synthesis.

Synthesized 15,6 s of speech audio.

Figure 49 Screenshot of DW Avatar (three views)

It uses the same speech synthesis code library as the DW Speaker and DW Podcast Creator apps. The customized DW Brazilian and Urdu synthetic voices that were trained in the SELMA project are used in this application, in addition to plain X Azure and Google voices, in order to have a wider range of voices and languages.

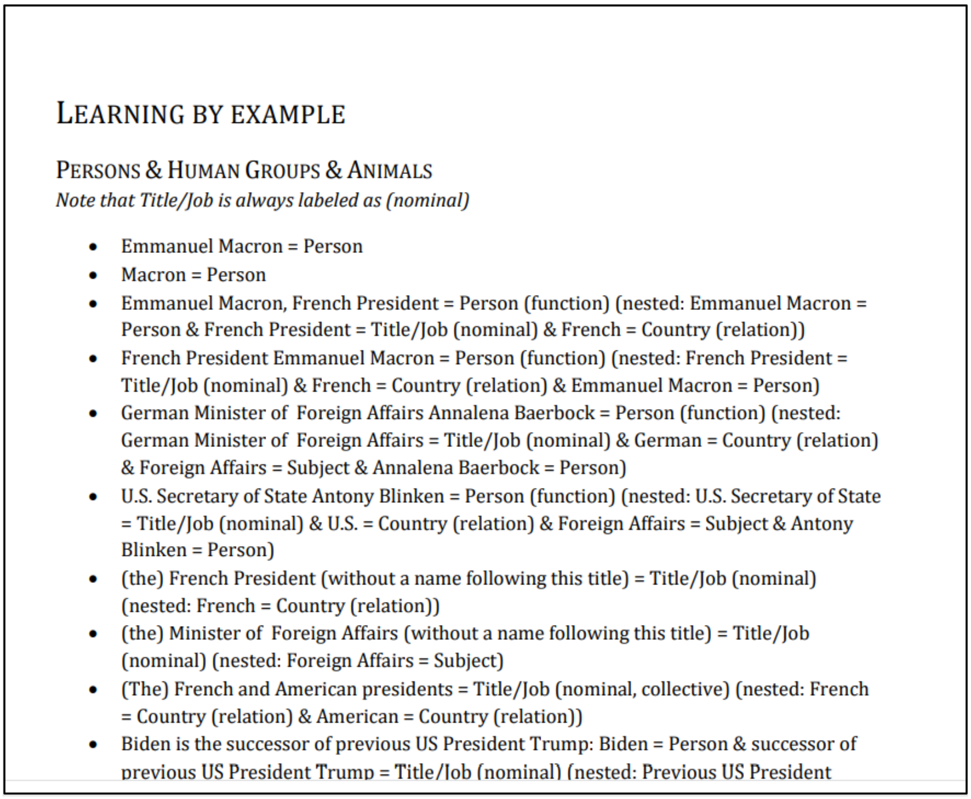
Conclusion:

The Avatar application shows how the SELMA speech synthesis code can easily be applied to other applications.

4.9 NER Component and Topic Detection

Year 1 determined the process for the technical as well as the user partners in terms of developing and training the named entity recognition component. It was agreed that certain languages will be targeted, and annotation should be done by native speakers. Training of the tool with previously created datasets was initiated. Annotation of some additional languages was started, including Ukrainian, Latvian and Russian. Training of DW project managers was started for Arabic.

In Year 2, the annotation process was streamlined and adapted, based on experiences on the first language set. The initial requirement to obtain 4,000 documents in each annotation language was revisited and reduced to 50-500, depending on the quality of the pre-annotation. This was necessary to keep the effort required for this task manageable and reasonable. A generic, multilingual annotator was built based on the first annotation sets, which allows for a pre-annotation of datasets in additional languages, thus reducing the need for such a high level of human annotation. The final human annotation level differs per language, for instance for Dutch, an annotation of 50 documents was sufficient for a very good result. Turkish, on the other hand, does not reveal such good results and needs more annotation. Deutsche Welle intensified the training and preparation and set up a detailed information package for the editors that will be involved in the annotation. This includes a special DW user guide for editors, with selected and sorted examples, to make the introduction as smooth as possible. Detailed feedback on the initial guidelines and the UI was provided to the annotation linguists. In this reporting period, the partners completed Latvian and Dutch, and started Russian, Ukrainian, Turkish, Arabic and Urdu.



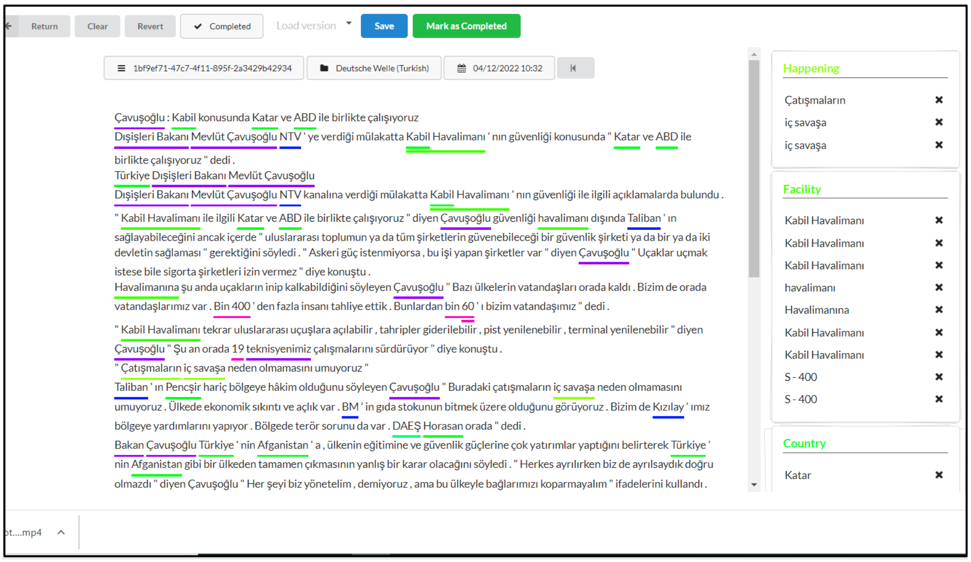
Learning by example

Persons & Human Groups & Animals

Note that Title/Job is always labeled as (nominal)

- Emmanuel Macron = Person
- Macron = Person
- Emmanuel Macron, French President = Person (function) (nested: Emmanuel Macron = Person & French President = Title/Job (nominal) & French = Country (relation))
- French President Emmanuel Macron = Person (function) (nested: French President = Title/Job (nominal) & French = Country (relation) & Emmanuel Macron = Person)
- German Minister of Foreign Affairs Annalena Baerbock = Person (function) (nested: German Minister of Foreign Affairs = Title/Job (nominal) & German = Country (relation) & Foreign Affairs = Subject & Annalena Baerbock = Person)
- U.S. Secretary of State Antony Blinken = Person (function) (nested: U.S. Secretary of State = Title/Job (nominal) & U.S. = Country (relation) & Foreign Affairs = Subject & Antony Blinken = Person)
- (the) French President (without a name following this title) = Title/Job (nominal) (nested: French = Country (relation))
- (the) Minister of Foreign Affairs (without a name following this title) = Title/Job (nominal) (nested: Foreign Affairs = Subject)
- (The) French and American presidents = Title/Job (nominal, collective) (nested: French = Country (relation) & American = Country (relation))
- Biden is the successor of previous US President Trump: Biden = Person & successor of previous US President Trump = Title/Job (nominal) (nested: Previous US President

Figure 50 Customized DW Guidelines for NER Annotation Editors



Çavuşoğlu : Kabil konusunda Katar ve ABD ile birlikte çalışıyoruz
Dışişleri Bakanı Mevlüt Çavuşoğlu NTV ' ye verdiği mülakatta Kabil Havalimanı ' nın güvenliği konusunda " Katar ve ABD ile birlikte çalışıyoruz " dedi .
Türkiye Dışişleri Bakanı Mevlüt Çavuşoğlu
Dışişleri Bakanı Mevlüt Çavuşoğlu NTV kanalına verdiği mülakatta Kabil Havalimanı ' nın güvenliği ile ilgili açıklamalarda bulundu .
" Kabil Havalimanı ile ilgili Katar ve ABD ile birlikte çalışıyoruz " diyen Çavuşoğlu güvenliği havalimanı dışında Taliban ' ın sağlayabileceğini ancak içerde " uluslararası toplumun ya da tüm şirketlerin güvenebileceği bir güvenlik şirketi ya da bir ya da iki devletin sağlaması " gerektiğini söyledi . " Askeri güç istenmiyorsa , bu işi yapan şirketler var " diyen Çavuşoğlu " Uçaklar uçmak istese bile sigorta şirketleri izin vermez " diye konuştu .
Havalimanına şu anda uçakların inip kalkabildiğini söyleyen Çavuşoğlu " Bazı ülkelerin vatandaşları orada kaldı . Bizim de orada vatandaşlarımız var . Bin 400 ' den fazla insanı tahliye ettik . Bunlardan bin 60 ' ı bizim vatandaşımız " dedi .
" Kabil Havalimanı tekrar uluslararası uçuşlara açılabilir , tahripler giderilebilir , pist yenilenebilir , terminal yenilenebilir " diyen Çavuşoğlu " Şu an orada 19 teknisyenimiz çalışmalarını sürdürüyor " diye konuştu .
" Çatışmaların iç savaşa neden olmamasını umuyoruz "
Taliban ' ın Pençşir hariç bölgeye hâkim olduğunu söyleyen Çavuşoğlu " Buradaki çatışmaların iç savaşa neden olmamasını umuyoruz . Ülkede ekonomik sıkıntı ve açlık var . BM ' in gıda stokunun bitmek üzere olduğunu görüyoruz . Bizim de Kızılay ' ımız bölgeye yardımlarını yapıyor . Bölgede terör sorunu da var . DAEŞ Horasan orada " dedi .
Bakan Çavuşoğlu Türkiye ' nin Afganistan ' a , ülkenin eğitimine ve güvenlik güçlerine çok yatırımlar yaptığını belirterek Türkiye ' nin Afganistan gibi bir ülkeden tamamen çıkmasının yanlış bir karar olacağını söyledi . " Herkes ayrılırken biz de ayrılsaydık doğru olmazdı " diyen Çavuşoğlu " Her şeyi biz yönetelim , demiyoruz , ama bu ülkeyle bağlarımızı koparmayalım " ifadelerini kullandı .

Happening
Çatışmaların ✖
iç savaşa ✖
iç savaşa ✖

Facility
Kabil Havalimanı ✖
Kabil Havalimanı ✖
Kabil Havalimanı ✖
havalimanı ✖
Havalimanına ✖
Kabil Havalimanı ✖
Kabil Havalimanı ✖
S - 400 ✖
S - 400 ✖

Country
Katar ✖

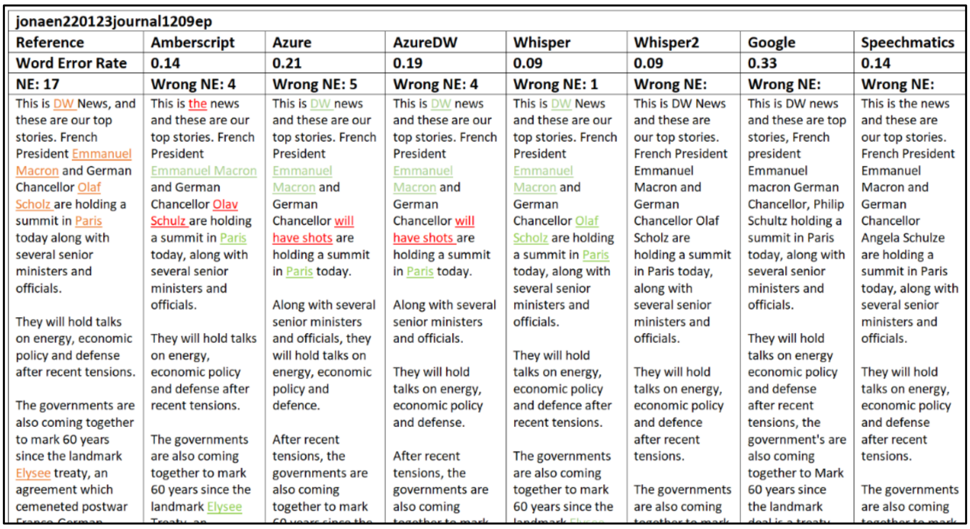
Figure 51 Sample of Turkish NER annotation

Year 3 focused on the use of NER and NEL in plain X and especially the Monitio platform and the NER API integration and testing through the DW Benchmarking tool. In addition, topic detection was also greatly improved.

Throughout the final project year, various iterations of the Monitio platform and the output and enhancements of the named entities extracted and displayed in the tool were assessed. User feedback was given to the developers, leading to further enhancements.

The NER API allows us to do an assessment of the level of named entity recognition when running an automated ASR evaluation on a specific transcription engine. More details on this integration can be found in the benchmarking section.

Below is an example of English-language comparative NER evaluation of 7 ASR engines, in which we analyzed several documents through the engines listed below. The NER analysis tool, hosted by Priberam and accessed by DW via API, provides a Word Error Rate (WER) for each text and each engine. This enables us to give users an overview of which ASR engines perform better in terms of named entities. This system is now in use at DW for assessment after an engine update or when a new engine becomes available.



jonaen220123journal1209ep							
Reference	Amberscript	Azure	AzureDW	Whisper	Whisper2	Google	Speechmatics
Word Error Rate	0.14	0.21	0.19	0.09	0.09	0.33	0.14
NE: 17	Wrong NE: 4	Wrong NE: 5	Wrong NE: 4	Wrong NE: 1	Wrong NE:	Wrong NE:	Wrong NE:
This is DW News, and these are our top stories. French President Emmanuel Macron and German Chancellor Olaf Scholz are holding a summit in Paris today along with several senior ministers and officials. They will hold talks on energy, economic policy and defense after recent tensions. The governments are also coming together to mark 60 years since the landmark Elysee treaty, an agreement which cemeneted postwar France-German	This is the news and these are our top stories. French President Emmanuel Macron and German Chancellor Olav Schulz are holding a summit in Paris today, along with several senior ministers and officials. They will hold talks on energy, economic policy and defense after recent tensions. The governments are also coming together to mark 60 years since the landmark Elysee Treaty, an	This is DW news and these are our top stories. French President Emmanuel Macron and German Chancellor will have shots are holding a summit in Paris today. Along with several senior ministers and officials, they will hold talks on energy, economic policy and defence. After recent tensions, the governments are also coming together to mark 60 years since the	This is DW news and these are our top stories. French President Emmanuel Macron and German Chancellor will have shots are holding a summit in Paris today. Along with several senior ministers and officials. They will hold talks on energy, economic policy and defense. After recent tensions, the governments are also coming together to mark	This is DW News and these are our top stories. French President Emmanuel Macron and German Chancellor Olaf Scholz are holding a summit in Paris today, along with several senior ministers and officials. They will hold talks on energy, economic policy and defence after recent tensions. The governments are also coming together to mark 60 years since the landmark Elysee	This is DW News and these are our top stories. French President Emmanuel Macron and German Chancellor Olaf Scholz are holding a summit in Paris today, along with several senior ministers and officials. They will hold talks on energy, economic policy and defence after recent tensions. The governments are also coming together to mark	This is DW news and these are top stories, French president Emmanuel macron German Chancellor, Philip Schultz holding a summit in Paris today, along with several senior ministers and officials. They will hold talks on energy economic policy and defense after recent tensions, the government's are also coming together to Mark 60 years since the landmark deal is a treaty	This is the news and these are our top stories. French President Emmanuel Macron and German Chancellor Angela Schulze are holding a summit in Paris today, along with several senior ministers and officials. They will hold talks on energy, economic policy and defense after recent tensions. The governments are also coming together to mark

Figure 52 *Shows an example of English-language comparative NER evaluation of 7 ASR engines*

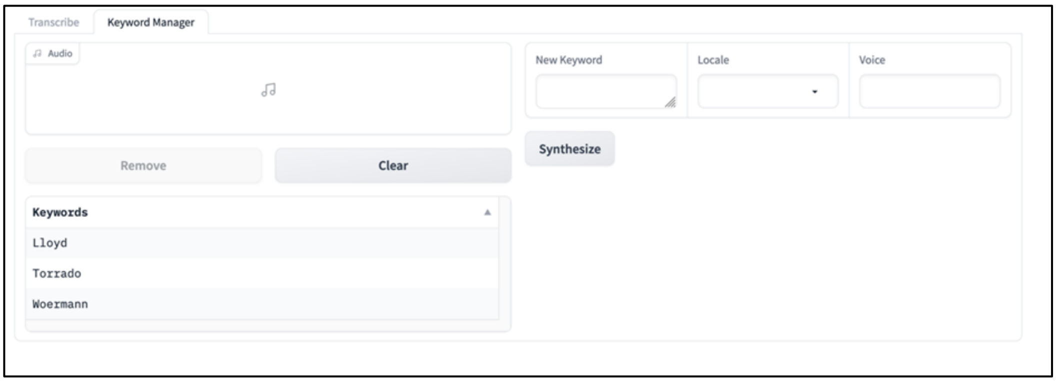
A final assessment of NER was done through the M-PHANTOM module providing a user-feedback system to assess and correct named entity output from ASR.

The standalone prototype M-PHANTOM allows users to correct named entities in a transcription and save them in a keyword manager. The transcription uses the Whisper3 model. To evaluate the functionality, we selected 4 videos in English and 1 video in French.

Each video is then transcribed, and users can correct the named entities in the transcript. The correction is then automatically identified and added to a database of keywords.

The transcription results were very accurate for all videos. Only three named entities needed to be corrected overall. In the example below, a French transcription, the named entity of "Woermann" was spelled wrongly.

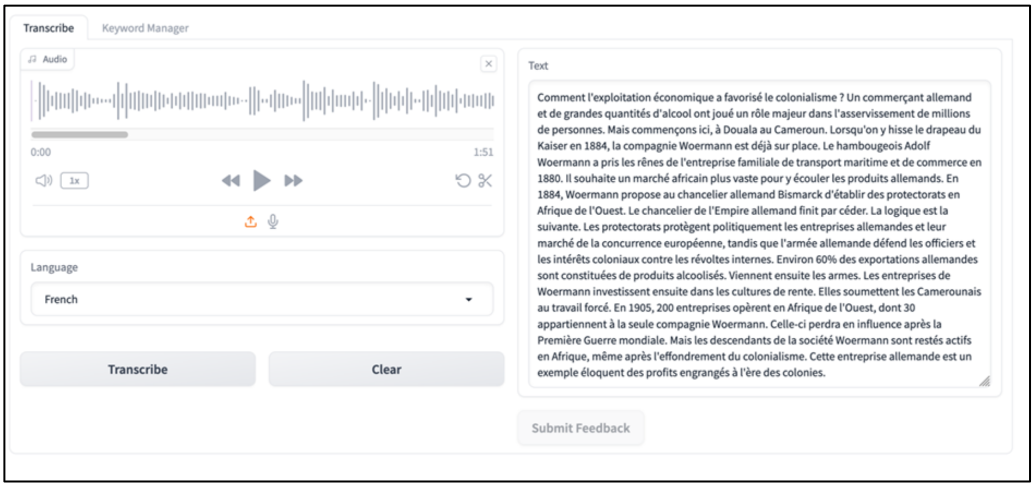
It was corrected, then added to the list of other keywords:



Keywords
Lloyd
Torrado
Woermann

***Figure* 53** *User evaluation in M-PHANTOM user correction model*

When the audio is reprocessed, the named entity is detected properly:



Language: French

Text

Comment l'exploitation économique a favorisé le colonialisme ? Un commerçant allemand et de grandes quantités d'alcool ont joué un rôle majeur dans l'asservissement de millions de personnes. Mais commençons ici, à Douala au Cameroun. Lorsqu'on y hisse le drapeau du Kaiser en 1884, la compagnie Woermann est déjà sur place. Le hambougeois Adolf Woermann a pris les rênes de l'entreprise familiale de transport maritime et de commerce en 1880. Il souhaite un marché africain plus vaste pour y écouler les produits allemands. En 1884, Woermann propose au chancelier allemand Bismarck d'établir des protectorats en Afrique de l'Ouest. Le chancelier de l'Empire allemand finit par céder. La logique est la suivante. Les protectorats protègent politiquement les entreprises allemandes et leur marché de la concurrence européenne, tandis que l'armée allemande défend les officiers et les intérêts coloniaux contre les révoltes internes. Environ 60% des exportations allemandes sont constituées de produits alcoolisés. Viennent ensuite les armes. Les entreprises de Woermann investissent ensuite dans les cultures de rente. Elles soumettent les Camerounais au travail forcé. En 1905, 200 entreprises opèrent en Afrique de l'Ouest, dont 30 appartiennent à la seule compagnie Woermann. Celle-ci perdra en influence après la Première Guerre mondiale. Mais les descendants de la société Woermann sont restés actifs en Afrique, même après l'effondrement du colonialisme. Cette entreprise allemande est un exemple éloquent des profits engrangés à l'ère des colonies.

***Figure* 54** *User feedback results in a named entity correction*

During the final reporting period, a new version of the classification module trained with more data was deployed in the Monitio platform. The previous version of the model showed unexpected behaviors when dealing with very small documents and that was corrected by introducing a second model on the pipeline to deal with those cases. The linguist team at Priberam evaluated a set of 666 documents in several languages comparing whether the set of topics attributed using the old model was better or worse than the last version and the result was that 214 documents had a better set of labels using the old and 436 were better using the latter, while 16 were indifferent. The next table shows the evaluation for each language.

Language	**Indifferent**	**Old is better**	**New is better**
ar	2	17	24
da		18	32
de	1	13	30
el		3	22
en	1	20	31
es		12	34
fa		8	33
fr		16	34
hu	2	17	30
it	1	12	33
ja			1

lv		13	27
pt	2	17	22
ru	5	15	27
tr	2	20	21
uk		13	35
	16	**214**	**436**

Table 3 *User topic detection evaluation*

Conclusion: Fortunately, we have experienced NER, NER and topic detection is improving and so are the tools to make it even better. Named entities are still one of the major obstacles in transcription, especially in a production environment. We aim at getting an extremely high accuracy for transcription, but names are often not recognized or misspelled. NER and NEL training, in particular for low-resourced languages, and named entity correction help alleviate these problems up to a certain degree. In this period, we evaluated both NER and the M-PHANTOM user feedback system in plain X, as well as topic detection improvements in Monitio.

4.10 Diarization Component

Deutsche Welle users evaluated diarization as implemented in the plain X platform, introducing the function to editors in the context of their productive work. Speaker labels were enabled in some ASR engines in the plain X demonstrator. See section 4.3.1 Diarization in this report.

The Fraunhofer Diarization Component is a standalone component and was mainly evaluated by Fraunhofer users and described in D3.7 and D3.8. DW users were able to assess the tool through a demo and compare the two systems.

The plain X diarization and its Fraunhofer counterpart work the same way, at least from the user's point of view. Both provide speaker labels, which are editable, timecodes, and ask for the number of expected speakers. The output quality is not perfect, it may occasionally add a wrong speaker label or fail to recognise a speaker shift, but overall users were very happy to have access to this feature.

The Fraunhofer diarization tool includes one major additional feature, viz. speaker recognition. It has a database of speaker names and attempts to recognize the voice and put a real speaker name on each voice when analyzing and annotating a video. In plain X and overall in SELMA components and joint platforms, we decided not to include speaker recognition in order not to infringe on any privacy rights. Thus, speaker recognition can lead to ethical issues. As the Fraunhofer speaker recognition module used only one type of content, namely German parliamentary debates as its only source, which is open data and concerns public speakers, such as politicians, this data could be used for this additional analysis. It would be of high value to the end users if this can be added to the NLP tools, but this would have to be under strict guidelines. It also requires careful monitoring, as it is likely that occasionally wrong labels will be assigned to speakers, leading to disinformation and wrongly contributing statements to speakers. The Fraunhofer tool has a good UI to correct such input, but, as said, close monitoring is important.

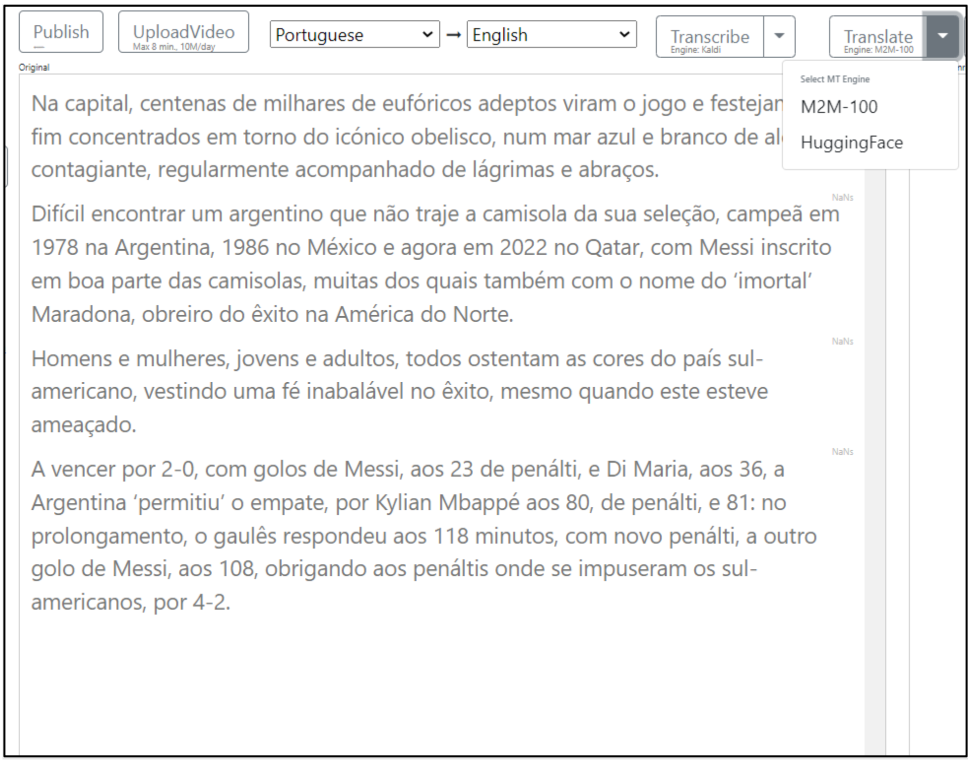
Conclusion: Diarization is part of an ASR module and dependent on whether the provider includes that feature. The modules that were assessed within SELMA (Fraunhofer's diarization tool and plain X using Azure and Amberscript) showed that this function works quite well and it is already in productive use at Deutsche Welle. Expanding the use of labeling to other engines is important, so more languages can be covered.

Bringing it to a next level by adding automated speaker recognition to diarization is very valuable and would be very useful, but this would require strict guidelines to protect privacy and avoid disinformation.

4.11 MT Components

Year 1 focused on setting the requirements for MT development and efforts, what is needed, what is already available, what should be integrated and what can be achieved within the project.

In terms of integration, we looked at which MT engines are/should be integrated into the different platforms in terms of efficiency and cost-effectiveness. UC0 is an open-source platform and costs should be kept to a minimum, especially if we want to make it available for wide-scale testing. M2M-100 and HuggingFace were selected as basic MT engines, as they cover a large number of languages, are fast enough and provide sufficient quality for functionality testing. The focus here is not on the translation quality, but on the workflow, processes and functionalities.



Na capital, centenas de milhares de eufóricos adeptos viram o jogo e festejar fim concentrados em torno do icónico obelisco, num mar azul e branco de al contagiante, regularmente acompanhado de lágrimas e abraços.

Difícil encontrar um argentino que não traje a camisola da sua seleção, campeã em 1978 na Argentina, 1986 no México e agora em 2022 no Qatar, com Messi inscrito em boa parte das camisolas, muitas dos quais também com o nome do 'imortal' Maradona, obreiro do êxito na América do Norte.

Homens e mulheres, jovens e adultos, todos ostentam as cores do país sul-americano, vestindo uma fé inabalável no êxito, mesmo quando este esteve ameaçado.

A vencer por 2-0, com golos de Messi, aos 23 de penálti, e Di Maria, aos 36, a Argentina 'permitiu' o empate, por Kylian Mbappé aos 80, de penálti, e 81: no prolongamento, o gaulês respondeu aos 118 minutos, com novo penálti, a outro golo de Messi, aos 108, obrigando aos penáltis onde se impuseram os sul-americanos, por 4-2.

Figure 55 *Generic MT engines in the SELMA OSS*

The plain X platform already offers a choice of MT engines, including DeepL, Google, Azure, and Facebook. Specific SELMA MT components will be added, in particular for direct speech-to-text and even speech-to-speech engines.

The Monitio platform is less focused on machine translation, but still includes it, to convert the content from different languages into the one(s) that the user has defined as the preferred language.

The overall focus of Machine Translation efforts in SELMA is on speech-to-text and speech-to-speech translation.

In **Year 2**, the speech-to-text translation module from the University of Avignon was added to the SELMA OSS and the plain X platforms. This allows a direct translation of speech within a video from English into French without going to a direct transcription.

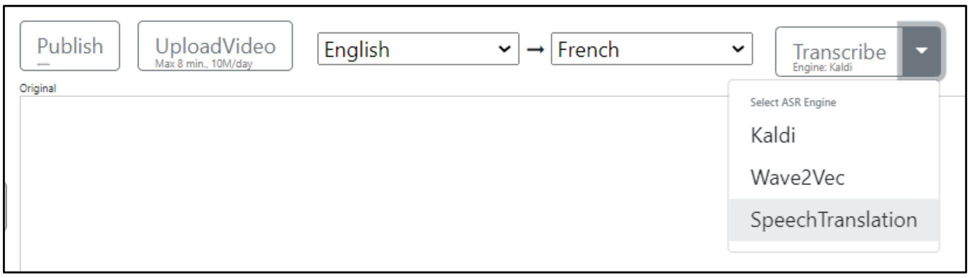


Figure 56 *ASR engines in the SELMA OSS*

This new module was evaluated by DW by testing the functionality in both platforms, to see if we actually get the French translation and how smooth the process is. At first, it worked only occasionally in plain X, due to integration issues. This feedback was forwarded to the developers and the feature was improved.

The next step was to compare the quality of the output of the traditional process, going through a transcription and then a translation, and the new process of going directly from speech to target language. Initial testing was done on videos from English into French translated text. This evaluation was ongoing in Y2.

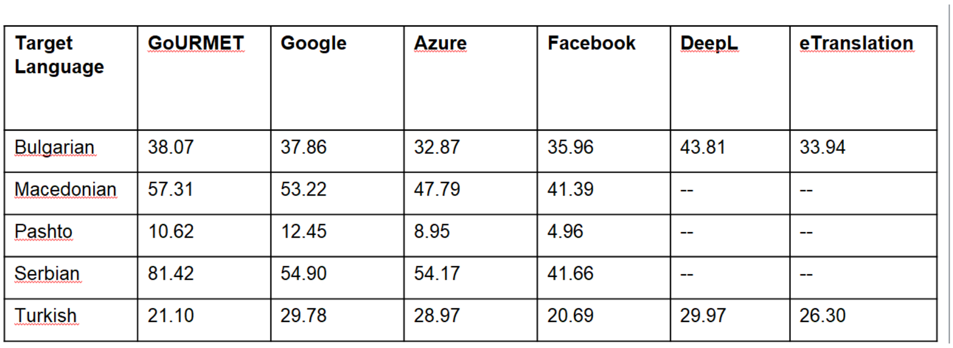
We also did an evaluation of some customized MT engines. Dockerized instances of the 32 GoURMET MT models for 16 languages (developed within the GoURMET H-2020 project) for some low-resourced languages were adapted, integrated and installed in the SELMA OSS to allow for an evaluation of the engines in the platform.

User evaluation was done at different levels:

- Texts were selected and ingested in batch into the SELMA OSS via API
- Back-to-back (reverse) translation for 16 languages was done
- A subsequent automated evaluation was done with BLEU scoring
- Reference scripts were created in different target languages
- BLEU scores were produced for languages which had a reference script
- Human evaluation was done of the best performing language pairs, based on the BLEU scores

More details of this process can be found in D5.2 section 4.9 on DW NLP Benchmarking.

The table below shows a comparative analysis with BLEU scores between the available engines for the best performing GoURMET-focused low-resource languages covered by DW from back-to-back translation evaluation using the SELMA OSS. It allows us to determine the best engine for a specific language pair.



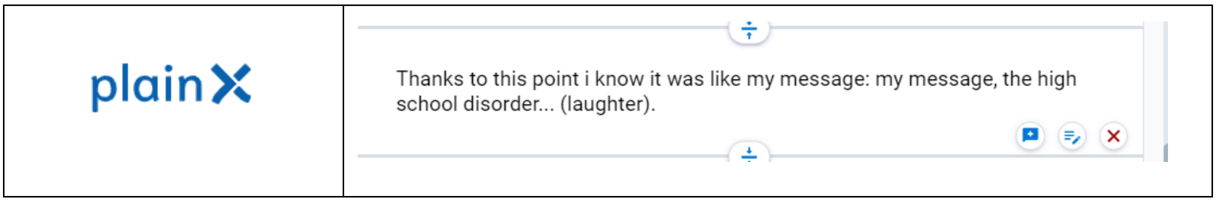
Target Language	GoURMET	Google	Azure	Facebook	DeepL	eTranslation
Bulgarian	38.07	37.86	32.87	35.96	43.81	33.94
Macedonian	57.31	53.22	47.79	41.39	--	--
Pashto	10.62	12.45	8.95	4.96	--	--
Serbian	81.42	54.90	54.17	41.66	--	--
Turkish	21.10	29.78	28.97	20.69	29.97	26.30

Table 4 *Comparative MT evaluation in SELMA OSS of selected engines*

In **Year 3**, we enhanced the benchmarking of the SELMA MT engines through the DW benchmarking tool. In this final year, the benchmarking tool itself was greatly improved and automated and the plain X API was integrated. This allowed us to do a fast and efficient benchmarking of the SELMA engines and compare them to output from other engines. More details can be found in this report’s section on benchmarking.

Furthermore, LIA's AST (automatic speech translation, i.e., speech-to-translated text) tool was re-evaluated after an update, including some major enhancements. The engine is now stable in plain X, works quite fast – at least the speed is comparable to the traditional combination of transcription followed by translation.

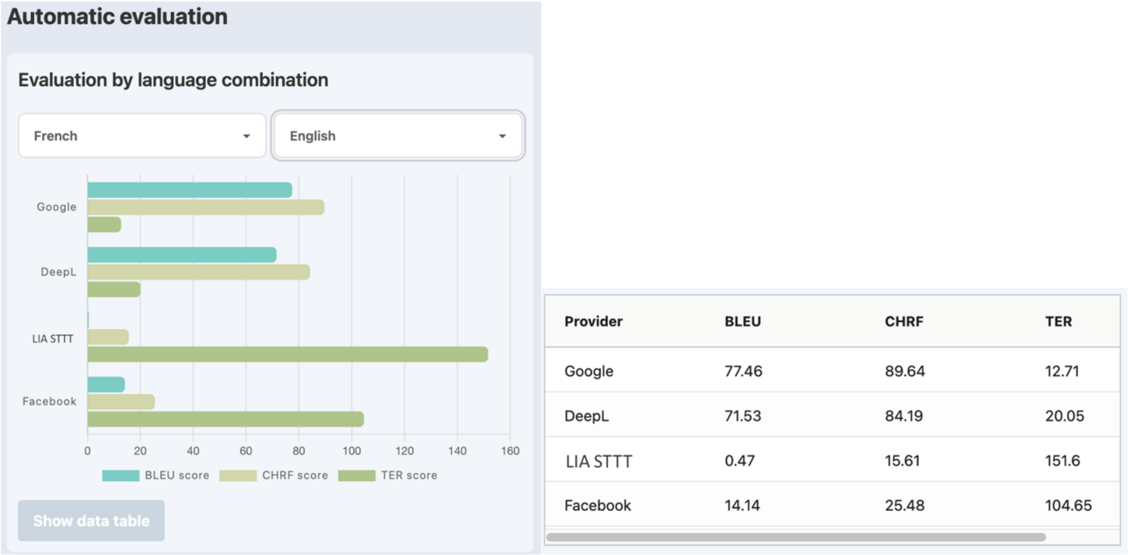
The conclusion is that this AST approach is definitely interesting for power MT users. Some languages do not have (good) ASR, and some aspects of the original speech can be retained. The AST output also includes audio descriptions, such as "laughter", "piano chords", "music", etc., which is a major step towards enhanced accessibility if such audio descriptions can be automated. They were not perfect - some were missing, others were incorrect, but it is a start and good to see this can be generated automatically.



Thanks to this point i know it was like my message: my message, the high school disorder... (laughter).

Figure 57 *Audio descriptions in the speech-to-translated-text module*

The output quality of the SELMA French-English AST could not (yet) compete with the traditional method of French transcription followed by a French-English machine translation if we can select the best separate engine from a combination of providers as is the case in plain X. However, if the quality gets better through training and further development and it is enhanced with certain aspects from the original speech, it may be a contender. In particular for languages that do not have any or good ASR, e.g. Tunesian, this could offer a very good solution. Definitely something that needs to be followed up.



Provider	BLEU	CHRF	TER
Google	77.46	89.64	12.71
DeepL	71.53	84.19	20.05
LIA STTT	0.47	15.61	151.6
Facebook	14.14	25.48	104.65

***Figure 58** Benchmarking of LIA's AST (automatic speech translation) engine*

LIA's speech-to-speech engine was only available as a stand-alone application and its evaluation was managed by LIA. The development status and quality output was not at the level required for implementation into one of the user tools. Development and progress on this component can be found in D3.7 (Final report on speech and natural language processing).

Other assessments in terms of MT from the user point of view concerned the overall MT workflow in the platforms, in particular in plain X, Monitio and OSS, as to ease of operation, speed, and accuracy.

In this final project year, the list of engines and providers in plain X was expanded, increasing the service offer towards the user. This is especially important for low-resourced languages. For instance, the LESAN engine was added for Amharic.

In Monitio, the translation capacity was increased, enabling MT from all DW languages into English, ensuring all DW published items are accessible to anyone within the DW network. This was a requirement for continued use of the Monitio platform at DW and it makes Monitio into a highly usable platform with transcription and translation for monitoring purposes.

MT as part of the SELMA OSS is also an essential part of the workflow, regardless of which engine is being applied. A simple, straightforward transcription-translation-voice-over workflow as released as open source opens up a lot of opportunities for certain user groups who do not require a full professional system. User assessment focused on the workflow offered by the OSS. Details are provided in the chapter dealing with SELMA OSS.

Conclusion: The progress we made in terms of machine translation focus on the experimental work in terms of STS (speech-to-speech) and AST (automatic speech translation, i.e., speech-to-translated-text) by LIA. These are definitely interesting and promising areas, but the quality of the output is still much too low to be used in a productive environment. Nevertheless, the opportunities are there, we are looking forward to more development here.

MT has also been a major part of our benchmarking efforts, also especially for the new engines that were added. Translation quality remains a discussion topic and benchmarking will play an increasingly important role.

4.12 ASR Components

Year 1 work included setting the requirements and priorities and determining what was available and what needed to be done. Both the University of Avignon already have some transcription tools, for instance for German, English, French, Arabic and Russian. LIA's main aim is to develop a speech-to-translated-text tool and a speech-to-speech tool and incorporate the ASR into these processes. The goal of the speech-to-speech translation component that LIA is working on is to transfer human-read news segments from one language to another, while keeping the original voice's expressivity.

FhG's goal is to apply its ASR to live broadcasting streams, to expand it to selected low-resourced languages, and work on enhancement modules such as punctuation.

Development and technical testing was the focus in the first period.

In **Year 2**, the user evaluation process was refined and started.

The user evaluation of speech-to-translated-text is covered in D5.2, section 4.7 on MT components.

We briefly describe the speech-to-speech module here. It has not yet been integrated in the user-oriented testing platforms plain X and OSS, but users were able to do a first assessment of the voices through a specific LIA evaluation application.

The evaluation's modus operandi is to play back corresponding pairs of news segments – one in the original language and one in the target language. Testers are asked to assess to what degree the original voice's expressivity has been transferred to the target language.

Two assessments are envisaged, both using the Likert scale. First, the *degree* of expressivity from 1 ('The target language shows no expressivity') to 5 ('The target language shows a human-like expressivity'). Second, the *fidelity* of expressivity, which assesses how truthfully the original's expressivity was transferred to the target language, from 1 ('the target segment's expressivity does not match the original') to 5 ('the target segment's expressivity matches the original').

During the SELMA User Day in October 2022, Deutsche Welle users did such user testing to specifically assess the expressivity from one language into another.

In **Year 3**, we did focused evaluation on the SELMA ASR modules, as well as evaluation of ASR as part of the benchmarking process, as described in the special section on benchmarking.

Specific evaluation of the separate components was done at the technology partner level and is described in deliverable D3.7 (Final report on speech and natural language processing). The diarization was also a key element of the evaluation of ASR results. This is further described in section 4.10.

LIA's AST (automatic speech translation, so speech-to-translated-text) module was compared with the traditional ASR + MT process and evaluated through use, evaluation and demos in the plain X platform (see previous section 4.11) and in the benchmarking tool (see section 4.15).

The benchmarking also includes an evaluation of the French SELMA ASR model.

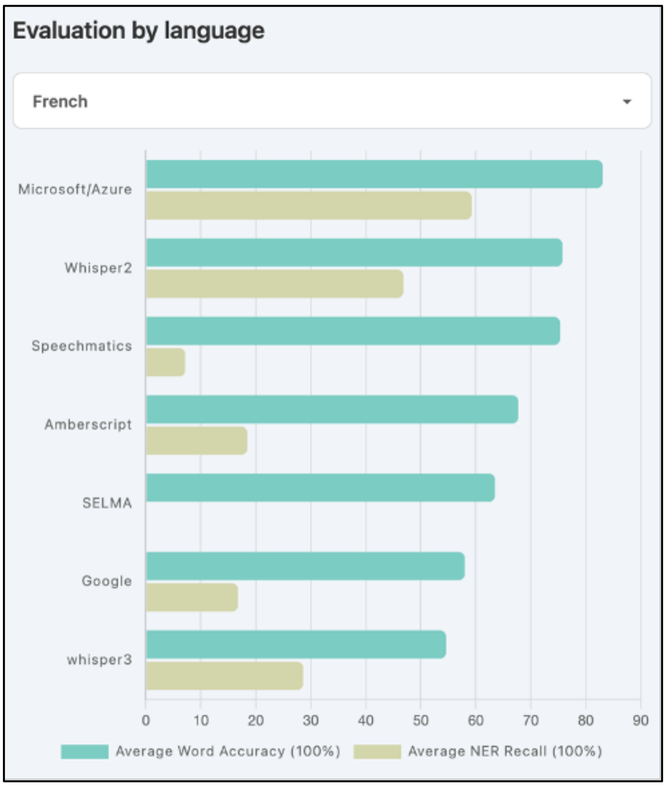
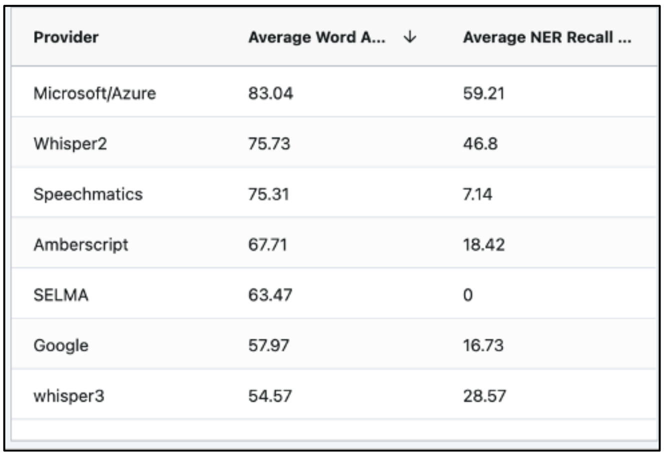


Figure 59 *French SELMA ASR benchmarking*

Below is the analysis of the benchmarking presented in numbers:



Provider	Average Word A... ↓	Average NER Recall ...
Microsoft/Azure	83.04	59.21
Whisper2	75.73	46.8
Speechmatics	75.31	7.14
Amberscript	67.71	18.42
SELMA	63.47	0
Google	57.97	16.73
whisper3	54.57	28.57

Figure 60 French SELMA ASR benchmarking in numbers

Deutsche Welle also provided training material to train some of the SELMA ASR engines for some low-resourced languages, in particular for Urdu, Bengali and Amharic, for which we supplied the technology partners with audio material and corresponding transcripts.

Conclusion: ASR has made progress and has been evaluated in particular in terms of the ASR enhancements of individual language components developed or improved in the project, including for Turkish, Russian, Bengali, Brazilian, and Tunesian dialect, as well as the AST (automatic speech translation) for French-English with audio descriptions, and improvements such as diarization in the demonstrators of Fraunhofer and plain X.

4.13 TTS Components

In **Year 1** we set the requirements for customizing synthetic voices for some of DW languages through a collaboration between DW and the University of Avignon, and selecting editorial departments and specific editors to be involved.

In **Year 2**, a total of eight voices of Brazilian DW journalists were cloned to develop the Brazilian text-to-speech component. This involved getting approval of the editors and collecting a dataset with audio and scripts with voices of the selected editors. These were collected by DW and handed over to LIA, who trained synthetic voices using that dataset.

On several iterations, the voices were assessed by the DW team and feedback was provided to the developers in terms of fluency, pronunciation accuracy and natural sound, including on interruptions in the output for certain voices, background noises (due to training from real content), robotic sounds, etc.. The assessment was repeated with new versions. The customized voices were integrated in the OSS, in plain X and in the podcasting application. Screenshots are included in the sections on the OSS and plain X platforms..

In **Year 3**, assessment of the Brazilian and Urdu synthetic voices was continued. In 2022, we aggregated 100h of audio news material in Brazilian Portuguese to model 9 TTS voices, corresponding to the voices of 9 DW colleagues who contributed to the training material. The latter ranged from 3 hours to 23 hours per contributor. The evaluation showed that the generated TTS audio contained music artifacts, hinting at insufficient segmentation of speech and music segments during the training phase. As a consequence, the voices were re-trained in 2023, thus producing much cleaner TTS output. The produced speech was still not perfect though, which led us to conclude that even more care needs to be put into cleaning up the training material.

The process was repeated with 10h of audio news material in Urdu at the end of 2023. This time, learning from the Brazilian voices, we placed a high emphasis on the separation of speech and music while preparing the training material. The resulting voice, 'Afsar', showed itself to be free from music artifacts. The voice is now being tested in a news production pilot project. First feedback from native listeners revealed that a few vowels, mostly decorated with diacritics are still being mispronounced, which lead us to deduce that the voice needs to be re-trained with more material, increasing the number of training hours from 10 to 20.

Conclusion: Text-to-speech technologies are on the rise and will undoubtedly be common practice in the near future for many applications. The voices are becoming so good and options for customization make them usable for publication. Deutsche Welle, and many other broadcasters, are ready to make them part of their production process. Training and customizing the Urdu and Brazilian voices have been a good exercise for the user partner team, learning that the technology works, but also where we need to step up to get the quality we stand for.

4.14 User Scenario Evaluation

In **Year 1**, we defined the use cases and the user scenarios. D1.1 - Use Case Description and Requirements has identified 22 scenarios, which are part of our evaluation effort. This relates to functionality testing (at platform level) with a specific purpose, namely that of each of the scenarios. As stated in the Use Case Description, the scenarios are functional areas identified as being relevant to SELMA and based on the personae and workflow descriptions as defined during the requirements process.

Evaluation is aimed at assessing and measuring their usability, accuracy and improvement over time.

In Year 2, all scenarios were active, except for

- Scenario 5: Generate Breaking News Alert
- Scenario 15: Highlight Item
- Scenario 17: Administer System
- Scenario 22: Apply Corrections

Table 5 below (User Scenarios) shows the final status in **Year 3,** listing the targeted scenarios, and provides a brief description for each and the focus of the evaluation. The last column indicates if evaluation of the particular scenario was done.

By the end of the SELMA project, we have implemented and evaluated all scenarios but one, viz, highlighting an item for team members in Monitio, which the consortium decided not to pursue. The highlight scenario as such was abolished, as its purpose was reached by other functions, such as saving items, saving and sharing a view, or using the send report function.

User Scenarios in Detail

#	Scenario ID	Scenario Description	Focus Evaluation	Evaluation Y3
1	Monitor Sources SEL-Sc-001	The user and language team specifies the input source(s) they wish to monitor through the system.	Functionality	Y
2	Ingest Media Item SEL-Sc-002	The system ingests media items from the sources.	Functionality	Y
3	Select Media Item SEL-Sc-003	The system selects media items and shows them to the user based on specific preferences.	Functionality	Y
4	Detect and Link Entity SEL-Sc-004	The system detects an entity and links it to other media items or clusters from the sources being monitored based on preferences specified by the user.	Functionality, Accuracy, Gradual Improvement	Y
5	Generate Breaking News Alert SEL-Sc-005	The system generates breaking news alerts based on the preferences set by the user.	Functionality, Relevance	Y
6	Create Transcription SEL-Sc-006	The system creates a transcription for an individual AV media Item.	Functionality, Accuracy, Gradual Improvement	Y
7	Create Translation SEL-Sc-007	The system creates a translation for an individual AV media item.	Functionality, Accuracy, Gradual Improvement	Y
8	View Cluster/ Entity SEL-Sc-008	The user views the details of a cluster and/or an entity.	Functionality	Y
9	View Individual Media Item	The user views an individual media item in relation to a cluster or entity.	Functionality	Y

	SEL-Sc-009			
10	Select Preferences SEL-Sc-0010	The user sets their preferences in the system.	Functionality, Usefulness	Y
11	Conduct Search SEL-Sc-0011	The user can search for an item in the system.	Functionality, Relevance of results	Y
12	Save Cluster / Individual Media Item SEL-Sc-0012	The user can save a cluster or an individual media item in the system where it is stored for more than a predefined set of time.	Functionality	Y
13	Remove Item SEL-Sc-0013	The user can remove an individual item and/or a cluster (with all its associated media items) from their view.	Functionality	Y
14	Train System SEL-Sc-0014	The user can train the system in relation to the cluster generation.	Functionality, Accuracy, Gradual Improvement	Y
15	Highlight Item SEL-Sc-0015	The user can highlight an item to make it visible to other members of the user's team.	Functionality	N*
16	Generate Trend Analysis SEL-Sc-0016	The system carries out a trend analysis and presents the results to the user.	Functionality, Accuracy, Usefulness, Gradual Improvement	Y
17	Administer System SEL-Sc-0017	The System Administrator carries out various activities to administer the system.	Functionality	Y
18	Group Media Items into Clusters SEL-Sc-0018	The system clusters media items based on the preferences set by the user.	Functionality, Relevance and Accuracy, Usefulness, Gradual Improvement	Y
19	Generate Summary SEL-Sc-0019	The system generates a summary for each media item.	Functionality, Relevance and Accuracy,	Y

			Usefulness, Gradual Improvement	
20	Generate Voice-Over SEL-Sc-0020	The system generates a voice-over for a transcription and/or translation of a media item on demand.	Functionality, Accuracy and Expressiveness, Gradual Improvement	Y
21	Edit Transcription/ Translation SEL-Sc-0021	The user can edit and correct the transcription and the translation. It is possible for 2 users to edit a transcription/translation simultaneously.	Functionality, Ease of use	Y
22	Apply Corrections SEL-Sc-0022	The system applies the corrections made by the user to the rest of the single media item or its cluster as defined by the user.	Functionality, Accuracy, Gradual Improvement	Y

Table 5 *User Scenarios in Detail*

*Not pursued, as it is covered by other functionalities

Conclusion: We implemented and enhanced all scenarios – referring to actions in the use cases – as foreseen, except for one, highlighting an item, which was deemed unnecessary, as we have other options. This shows that the use case demonstrators are very rich in terms of actions.

4.15 DW NLP Benchmarking

DW puts great effort in performing in-house benchmarking (BM) for the major NLP processes through direct use by the users, i.e., ASR (automated speech recognition) and MT (machine translation). It is important to have full control over which languages we can evaluate when and this within a short timeframe.

In **Year 1**, we established the benchmarking procedure and prepared the evaluation material.

This evaluation was started and is currently in process for all 32 DW languages and consists of both a human and an automated evaluation. A dataset was selected to serve as a baseline. For ASR evaluation, we use up to three videos per target language, the videos are selected from the DW archiving system that already has an editorial script. This (manu)script is checked by an editor from the corresponding language department to assess accuracy. The automated evaluation is then performed by calculating the Word Error Rate (WER). We aim to use 3 videos for the benchmarking of each target language.

For the evaluation of MT, we use five videos that have been preselected in English and German, with the corresponding transcripts. We then request each editorial department to provide a reference text in the target language, sentence-aligned, for each of the videos, from either the English or German source transcript. Once we are provided with the reference texts and have obtained the output texts from each MT engine, we can do the actual human and automated assessment. The BLEU score was selected as a rating score.

The automated evaluation is supplemented by a user assessment, i.e., an evaluation of the quality of the transcription or machine translation output -- for all engines available to the DW team for the language (pair) being assessed -- by a native speaker proficient in the source language (as well as the target language in the case of MT). A rating is made for different aspects, including translation accuracy, punctuation and capitalization, fluency, completeness. Human evaluation is done by means of user questionnaires and Likert ratings of 1-5 on user satisfaction.

An initial partial dataset was created to cover a few languages with which the first evaluations were done.

In **Year 2**, the reference dataset was expanded to more languages. Editorial departments were involved to provide the reference texts, i,e., (1) Check the accuracy of the transcript of the video selected in their target language against the audio content and (2) Provide a human translation for the English or German text of the five selected reference videos. Delivery of the edited content depends on the availability of the editors in the different language departments. Editors were asked to provide one ASR file and the first MT reference text as a priority, so the evaluation for their target language could be started. In this reporting year, editorial reference material was produced for Kiswahili, Serbian, Pashto, French, Spanish, Portuguese, Arabic, Indonesian, Urdu, Chinese, Bengali, Russian, Turkish, Ukrainian, Macedonian, Persian, Polish.

The list of NLP tools was expanded to include:

- 5 for ASR: Amberscript, Google, Azure, Speechmatics, and most recently Whisper
- 6 for MT: Google, Azure, Facebook, DeepL, eTranslation, GoURMET

We also asked the editors (native speakers) to assess the output of all available MT output and provide their feedback using the user questionnaire which was set up for this purpose, providing a satisfaction rating of 1 to 5. For some languages this meant evaluating up to 40 MT output texts, e.g. 5 texts in 4 tools (for instance Google, Azure, DeepL, Facebook) from two source languages (English and German).

We also did back-to-back translations for all languages, so that we can do a comparative analysis even if no reference text is available (yet). This provided us with a basic evaluation of MT output of the different MT engines. It compares the English input and output text after an automated translation from English into the target language and then translating that output text back into English using the same provider, for instance Azure MT into English and translating that output back into English with Azure. This is also shown in Figure 1.

The automated evaluation process was refined, and it was decided to expand the MT automated ratings to three metrics for each language pair: BLEU, chrF and TER (Translation Error Rate). This provides a more reliable evaluation output.

To support the automatic evaluation of both ASR and MT tasks, the development of a web application was started and a first version is ready and in use. This tool allows users to upload both reference and output texts, calculate the metrics and store results into a database. The user can also obtain MT engine output from a source text directly and perform the automatic evaluation for MT if a reference text is available.

In addition to ASR and MT, we did some benchmarking on speech synthesis. This is explained in D5.2 section 4.9.

In **Year 3**, the Benchmarking (BM) web application was enhanced. The dataset was expanded, user evaluations were done and the benchmarking process was automated.

The procedure for the automated assessment enables a very fast and consistent evaluation, which is essential in case of updates or newly available engines.

We also explore ways of providing automated support for human evaluation of ASR and MT, with an analysis of the user input coming from the user questionnaires, making the process much faster and reliable.

We further expanded the reference dataset and additional editors were involved, native speakers of the target language.

The main work was the automation of the process and the development work that this entailed.

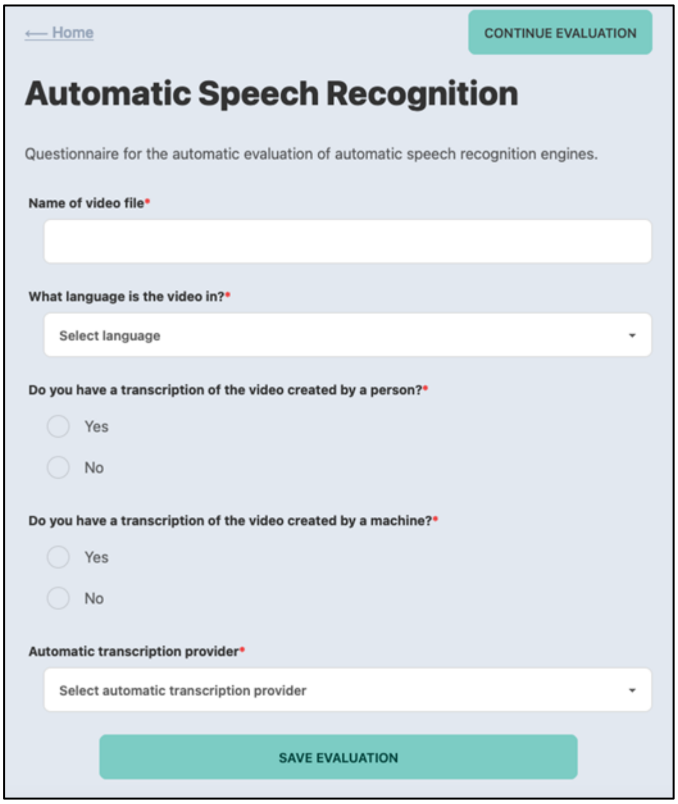
The BM process is now well established and structured with most of the NLP processes at least partially automated.

Below we describe the current BM process.

Automated Speech Recognition

To evaluate ASR engines, we use two metrics: the Word Error Rate (WER) and the Named Entity recall (NER).
The WER counts the numbers of insertions, deletions and substitutions in the hypothesis transcript compared to the gold standard.
In the Benchmarking platform, the WER is calculated using the open-source library jiwer.
The NER recall indicates how many named entities have been correctly transcribed in the hypothesis transcript. This metric uses the MONITIO API for named entity recognition.

In the SELMA BM platform, the user needs to upload the gold transcription and the hypothesis transcription.



Automatic Speech Recognition

Questionnaire for the automatic evaluation of automatic speech recognition engines.

Name of video file*

What language is the video in?*

Select language

Do you have a transcription of the video created by a person?*

○ Yes

○ No

Do you have a transcription of the video created by a machine?*

○ Yes

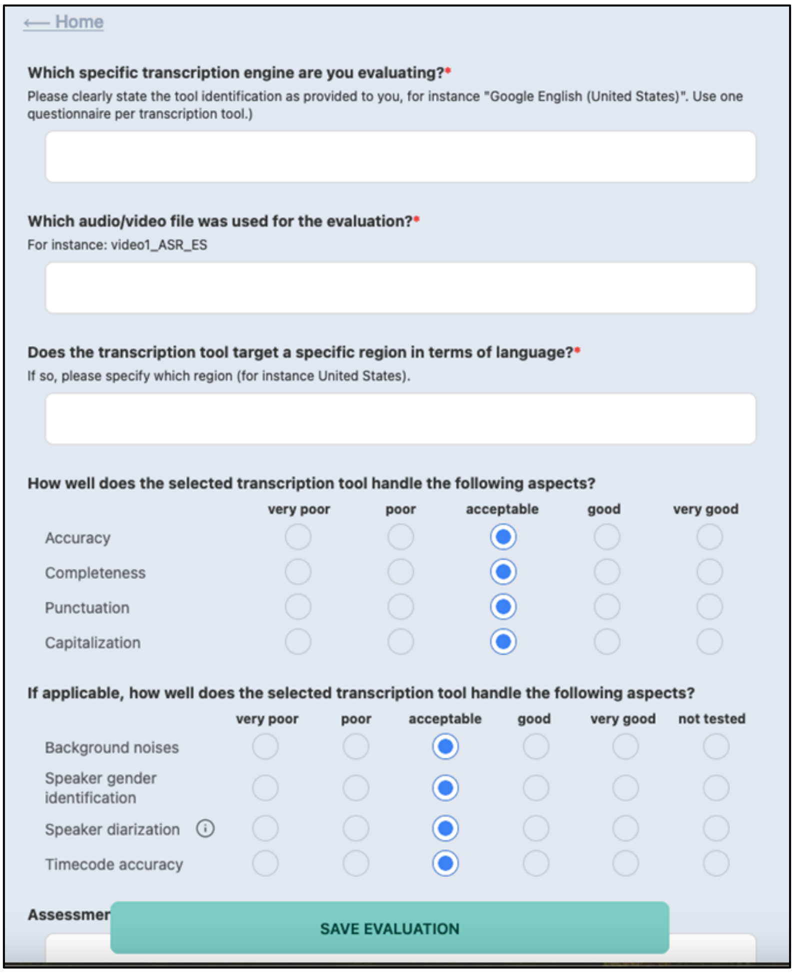
○ No

Automatic transcription provider*

Select automatic transcription provider

***Figure* 61** *ASR BM Upload Screen*

Users can also evaluate the hypothesis transcript manually and review the quality of the engine in the platform.



Which specific transcription engine are you evaluating?*
Please clearly state the tool identification as provided to you, for instance "Google English (United States)". Use one questionnaire per transcription tool.)

Which audio/video file was used for the evaluation?*
For instance: video1_ASR_ES

Does the transcription tool target a specific region in terms of language?*
If so, please specify which region (for instance United States).

How well does the selected transcription tool handle the following aspects?

	very poor	poor	acceptable	good	very good
Accuracy	○	○	●	○	○
Completeness	○	○	●	○	○
Punctuation	○	○	●	○	○
Capitalization	○	○	●	○	○

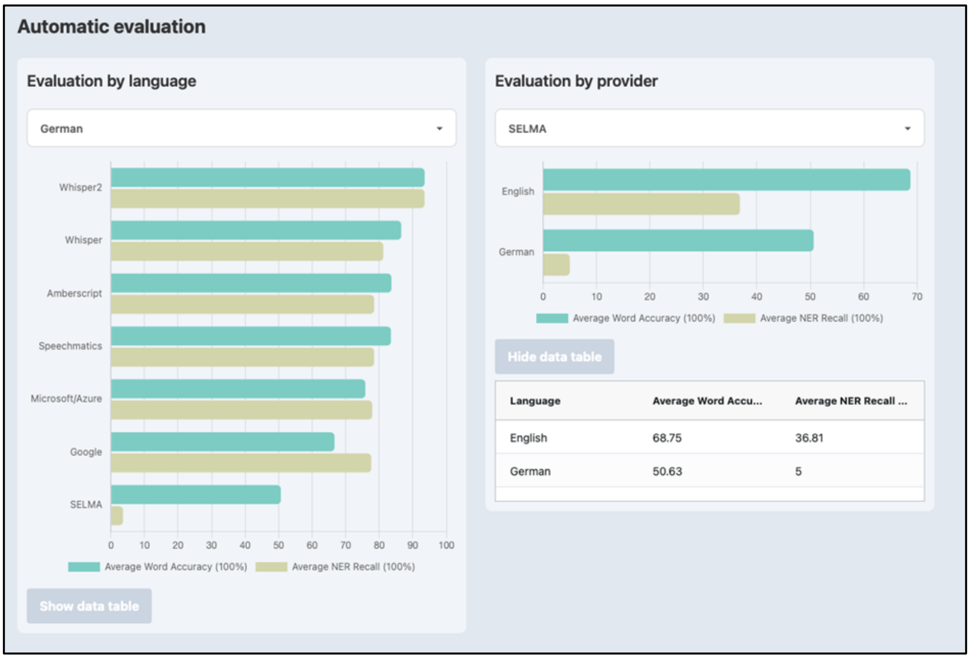
If applicable, how well does the selected transcription tool handle the following aspects?

	very poor	poor	acceptable	good	very good	not tested
Background noises	○	○	●	○	○	○
Speaker gender identification	○	○	●	○	○	○
Speaker diarization ⓘ	○	○	●	○	○	○
Timecode accuracy	○	○	●	○	○	○

Assessmer

***Figure* 62** *Human Evaluation for ASR BM*

The results are directly displayed in the platform. Users can choose to visualize the results of the automatic or manual evaluation based on 1) the provider, or 2) the language.



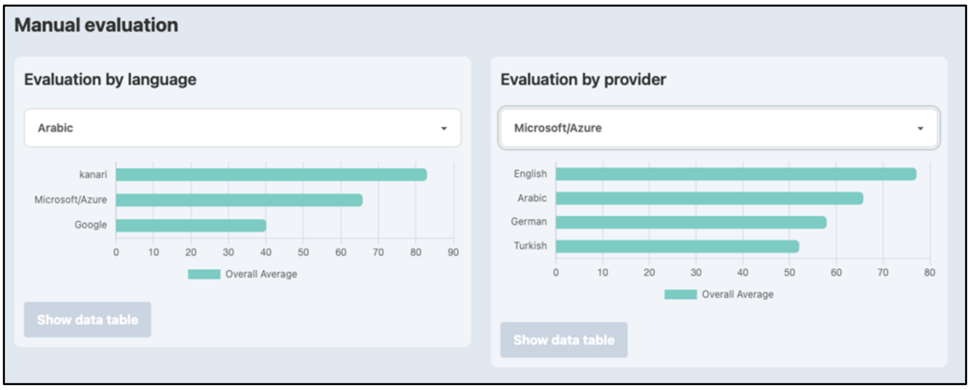
Evaluation by language: German

Evaluation by provider: SELMA

Language	Average Word Accu...	Average NER Recall ...
English	68.75	36.81
German	50.63	5

***Figure* 63** *Display of Automated BM Results for ASR*

Below is a screenshot of how the human ASR evaluation results are shown in the user interface for benchmarking.



Evaluation by language: Arabic

Evaluation by provider: Microsoft/Azure

***Figure* 64** *Display of Human BM Results for ASR*

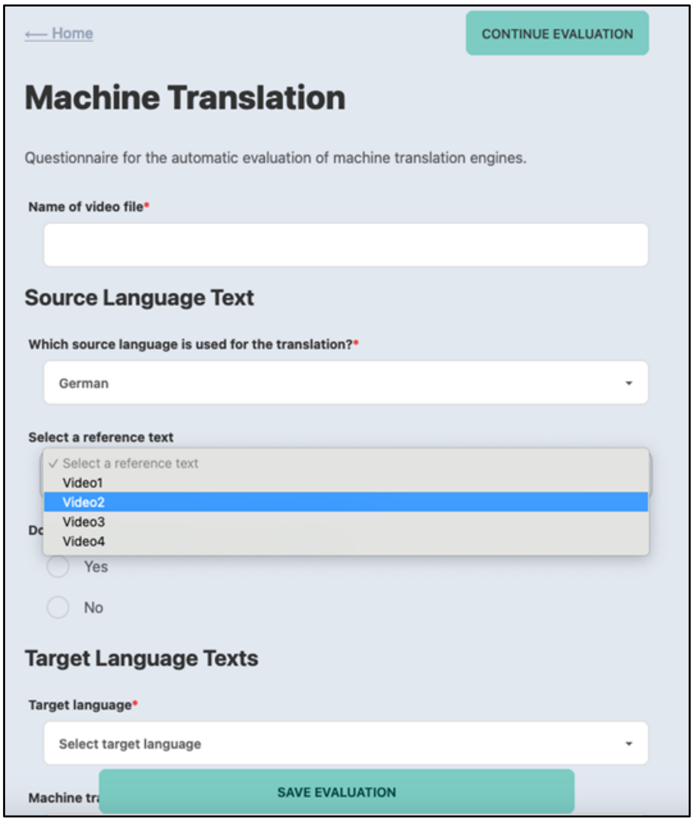
Machine Translation

To evaluate translation engines, we use 3 metrics: the BLEU score, the character-level F-score (chrF) and the Transcription Error Rate (TER). These scores are calculated independently.

The BLEU score measures how similar the hypothesis translation is to the gold text based on the sequence of words. The chrF is similar to the BLEU score but is based on the sequence of characters between the hypothesis and the ground truth. The TER score measures how much a human would have to edit the hypothesis text to match the reference.
In the benchmarking platform, these metrics are implemented using the SacreBLEU library (https://github.com/mjpost/sacrebleu).

In the SELMA BM platform, the user can choose to upload the reference text of their choice or choose a reference text from the available ones. The user can then upload the output translation in the chosen target language manually or request the translation from one of the providers available through the platform.

Below is an image of the Benchmarking Tool with MT input screen with source language information.



⟵ Home

CONTINUE EVALUATION

Machine Translation

Questionnaire for the automatic evaluation of machine translation engines.

Name of video file*

Source Language Text

Which source language is used for the translation?*

German

Select a reference text

✓ Select a reference text
Video1
Video2
Video3
Video4

Dc

○ Yes

○ No

Target Language Texts

Target language*

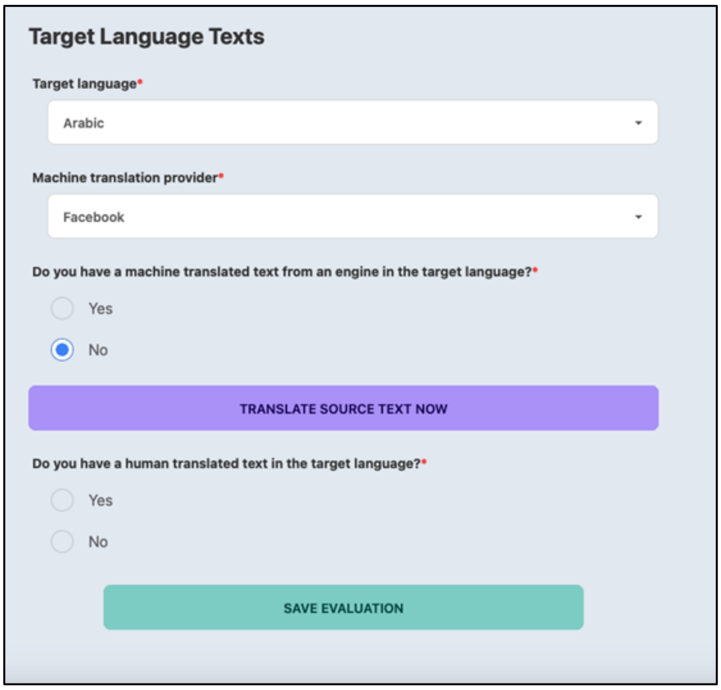
Select target language

Machine tr

SAVE EVALUATION

***Figure* 65** *MT BM Upload Screen – source language*

And this is a view of the input screen for the target language.



Target Language Texts

Target language*

Arabic

Machine translation provider*

Facebook

Do you have a machine translated text from an engine in the target language?*

○ Yes

● No

Do you have a human translated text in the target language?*

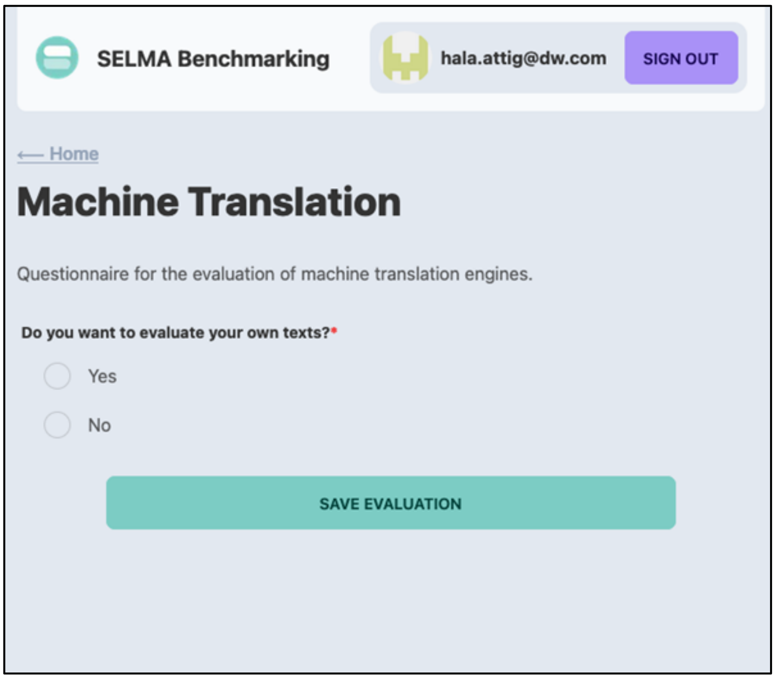
○ Yes

○ No

***Figure* 66** *MT BM Upload Screen – target language*

Users can also evaluate the quality of a translation manually in the platform. They can choose to evaluate a translation from an external source (i.e., a text locally saved) or choose one of the evaluations already present in the system. The text in the source language and the text in the target language from an engine will then appear side by side in the platform, allowing the user to fill in the question while reading the translation.

Below we see the start input screen for human evaluation of machine translation output.



Machine Translation

Questionnaire for the evaluation of machine translation engines.

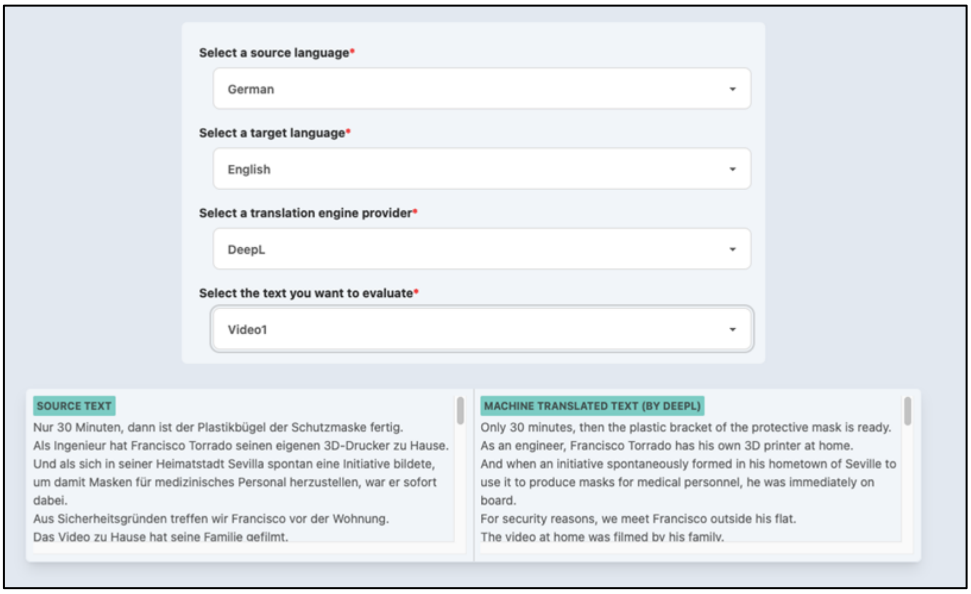
Do you want to evaluate your own texts?*

○ Yes

○ No

Figure 67 *Human Evaluation Questionnaire for MT BM*

Next the information that needs to be provided for human evaluation of MT output.



Select a source language*: German

Select a target language*: English

Select a translation engine provider*: DeepL

Select the text you want to evaluate*: Video1

SOURCE TEXT

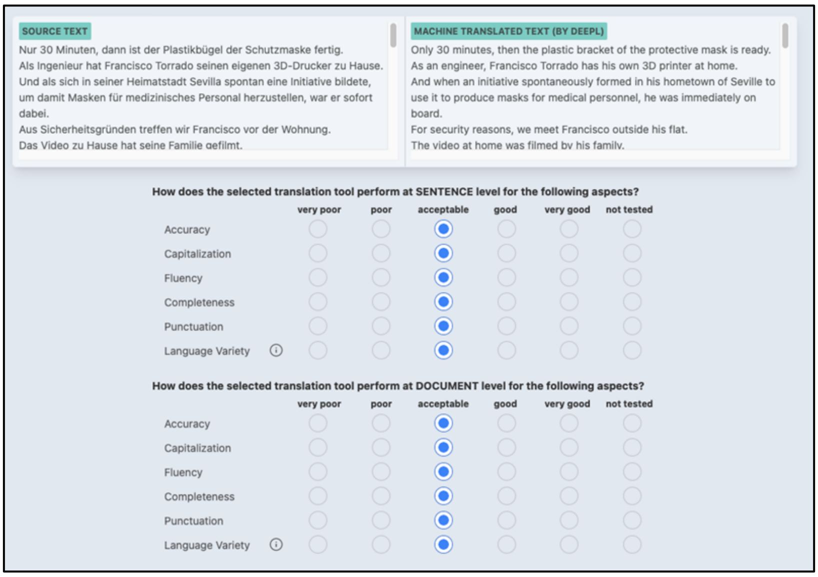
Nur 30 Minuten, dann ist der Plastikbügel der Schutzmaske fertig.
Als Ingenieur hat Francisco Torrado seinen eigenen 3D-Drucker zu Hause.
Und als sich in seiner Heimatstadt Sevilla spontan eine Initiative bildete, um damit Masken für medizinisches Personal herzustellen, war er sofort dabei.
Aus Sicherheitsgründen treffen wir Francisco vor der Wohnung.
Das Video zu Hause hat seine Familie gefilmt.

MACHINE TRANSLATED TEXT (BY DEEPL)

Only 30 minutes, then the plastic bracket of the protective mask is ready.
As an engineer, Francisco Torrado has his own 3D printer at home.
And when an initiative spontaneously formed in his hometown of Seville to use it to produce masks for medical personnel, he was immediately on board.
For security reasons, we meet Francisco outside his flat.
The video at home was filmed by his family.

Figure 68 *Human Evaluation for MT BM input*

The image below shows how the evaluator enters his assessment, using a ranking of different aspects of the translation.



SOURCE TEXT

Nur 30 Minuten, dann ist der Plastikbügel der Schutzmaske fertig.
Als Ingenieur hat Francisco Torrado seinen eigenen 3D-Drucker zu Hause.
Und als sich in seiner Heimatstadt Sevilla spontan eine Initiative bildete, um damit Masken für medizinisches Personal herzustellen, war er sofort dabei.
Aus Sicherheitsgründen treffen wir Francisco vor der Wohnung.
Das Video zu Hause hat seine Familie gefilmt.

MACHINE TRANSLATED TEXT (BY DEEPL)

Only 30 minutes, then the plastic bracket of the protective mask is ready.
As an engineer, Francisco Torrado has his own 3D printer at home.
And when an initiative spontaneously formed in his hometown of Seville to use it to produce masks for medical personnel, he was immediately on board.
For security reasons, we meet Francisco outside his flat.
The video at home was filmed by his family.

How does the selected translation tool perform at SENTENCE level for the following aspects?

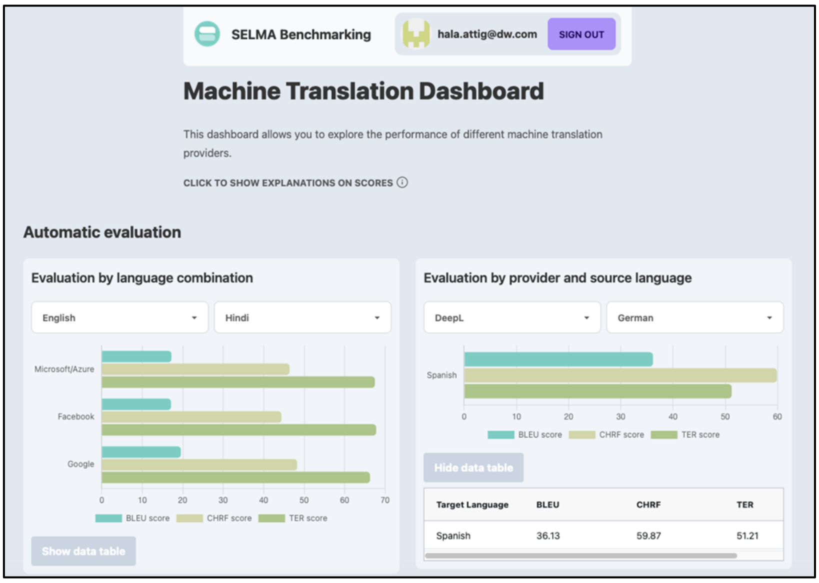
	very poor	poor	acceptable	good	very good	not tested
Accuracy	○	○	●	○	○	○
Capitalization	○	○	●	○	○	○
Fluency	○	○	●	○	○	○
Completeness	○	○	●	○	○	○
Punctuation	○	○	●	○	○	○
Language Variety ⓘ	○	○	●	○	○	○

How does the selected translation tool perform at DOCUMENT level for the following aspects?

	very poor	poor	acceptable	good	very good	not tested
Accuracy	○	○	●	○	○	○
Capitalization	○	○	●	○	○	○
Fluency	○	○	●	○	○	○
Completeness	○	○	●	○	○	○
Punctuation	○	○	●	○	○	○
Language Variety ⓘ	○	○	●	○	○	○

***Figure* 69** *Human Evaluation for MT BM - Ranking*

The results are displayed immediately in the platform. The user can choose to visualize the results of the automatic or manual evaluation based on 1) the language pair or 2) the source language and provider. Below we see how the results of automated MT evaluation is presented.



SELMA Benchmarking

hala.attig@dw.com

SIGN OUT

Machine Translation Dashboard

This dashboard allows you to explore the performance of different machine translation providers.

CLICK TO SHOW EXPLANATIONS ON SCORES ⓘ

Automatic evaluation

Evaluation by language combination

English ▾ Hindi ▾

Show data table

Evaluation by provider and source language

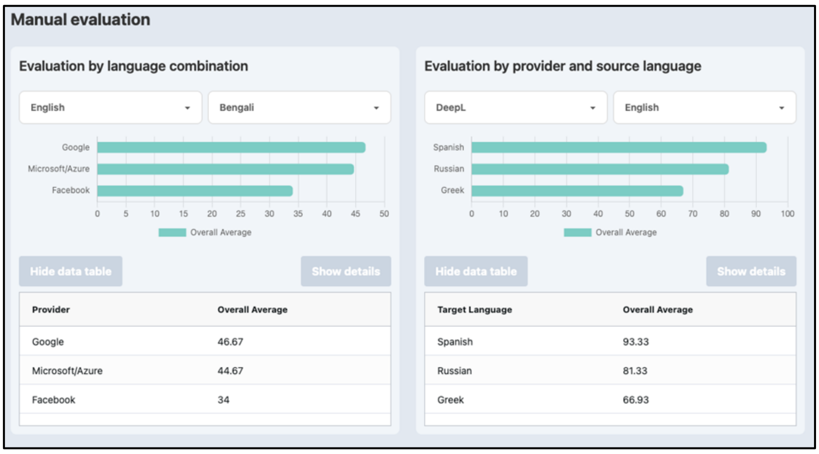
DeepL ▾ German ▾

Hide data table

Target Language	BLEU	CHRF	TER
Spanish	36.13	59.87	51.21

***Figure* 70** *Display of Automated BM Results for MT*

Next we show how the results of a human MT evaluation is presented.



Manual evaluation

Evaluation by language combination

Provider	Overall Average
Google	46.67
Microsoft/Azure	44.67
Facebook	34

Evaluation by provider and source language

Target Language	Overall Average
Spanish	93.33
Russian	81.33
Greek	66.93

***Figure* 71** *Display of Human BM Results for MT*

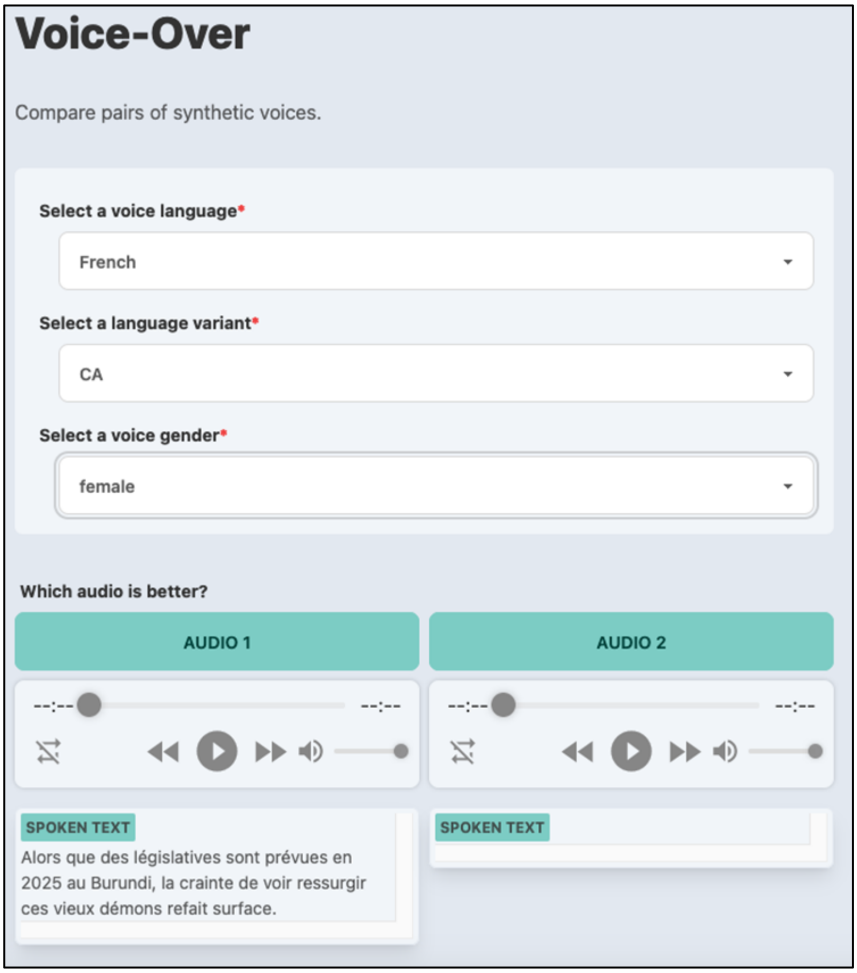
Voice-Over

There are no metrics to evaluate a synthetic voice. The evaluation needs to be done manually. For this, users are asked to listen to a voice and fill in a given form.

However, for high-resourced languages, this task becomes time-consuming and frustrating, as for one provider and one language, we can have more than 20 voices to evaluate. Although the evaluation of a single voice is important when developing synthetic voices, the use case for DW focused on answering the question: “which voice is the best for a certain language and variant?”. To answer this, we have decided to ask users not to evaluate a voice, but to rank them within the same language and variant. We also distinguish between female and male voices.

Once the user selects which language, variant and gender they want to evaluate, the system displays two audio files with the corresponding text. After playing the audio, the user then decides which voice is best between the two and clicks on it.

The next step is to complete the development to have this saved in the system, and then have it display two new voices to compare, until all voices have been compared at least once between each other. Each comparison is then saved individually, so the user can leave the evaluation at any time.



Voice-Over

Compare pairs of synthetic voices.

Select a voice language*

French

Select a language variant*

CA

Select a voice gender*

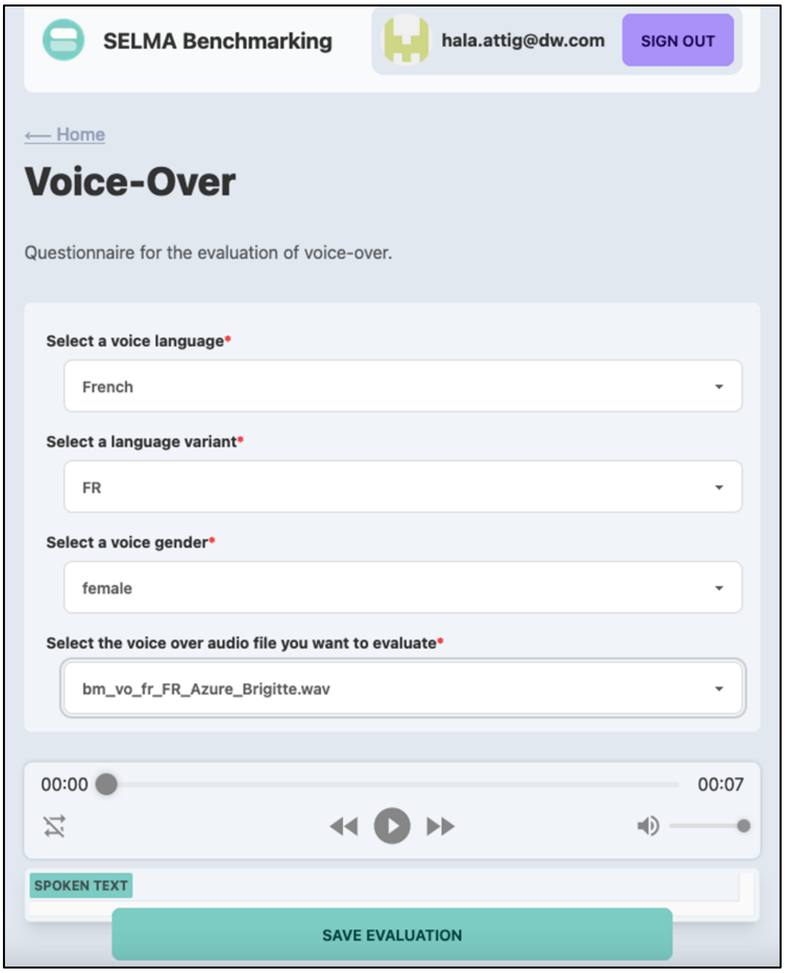
female

Which audio is better?

Alors que des législatives sont prévues en 2025 au Burundi, la crainte de voir ressurgir ces vieux démons refait surface.

Figure 72 Comparative Human VO BM

The manual evaluation of a single synthetic voice is also possible in the platform. The user needs to select the language, variant and gender and the system then displays the available audio files. Once an audio file selected, the user is able to listen to the synthetic voice while filling in the questionnaire.



SELMA Benchmarking

hala.attig@dw.com

Voice-Over

Questionnaire for the evaluation of voice-over.

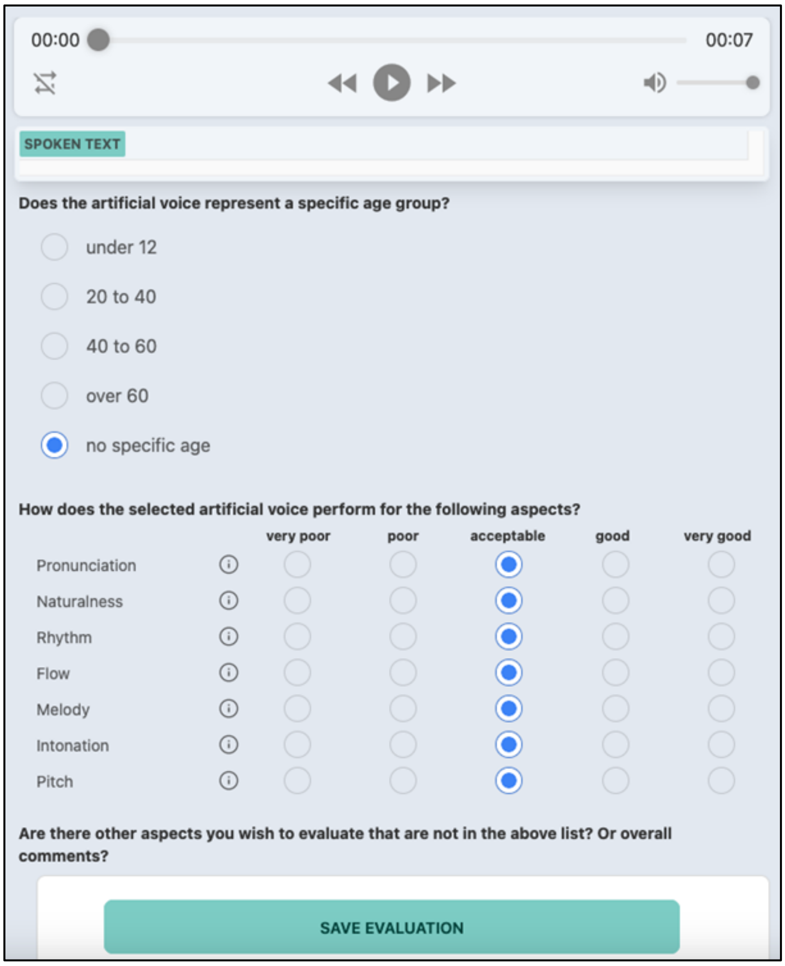
Select a voice language*
French

Select a language variant*
FR

Select a voice gender*
female

Select the voice over audio file you want to evaluate*
bm_vo_fr_FR_Azure_Brigitte.wav

***Figure** 73* Manual Human Assessment of a VO



Does the artificial voice represent a specific age group?

- ○ under 12
- ○ 20 to 40
- ○ 40 to 60
- ○ over 60
- ● no specific age

How does the selected artificial voice perform for the following aspects?

	very poor	poor	acceptable	good	very good
Pronunciation	○	○	●	○	○
Naturalness	○	○	●	○	○
Rhythm	○	○	●	○	○
Flow	○	○	●	○	○
Melody	○	○	●	○	○
Intonation	○	○	●	○	○
Pitch	○	○	●	○	○

Are there other aspects you wish to evaluate that are not in the above list? Or overall comments?

Figure 74 Manual Human Assessment of a VO – Input Screen

Conclusion:

The benchmarking tool and process has made tremendous progress over the past year and can now be used for fast and efficient automated evaluation of new or updated engines or to provide a new comparison of engines upon request. It will continue to be expanded with more reference texts in its database and is therefore a living system.

5. Timeline

We are in line with the estimated timeline that was set in D5.1 - Evaluation Plan.

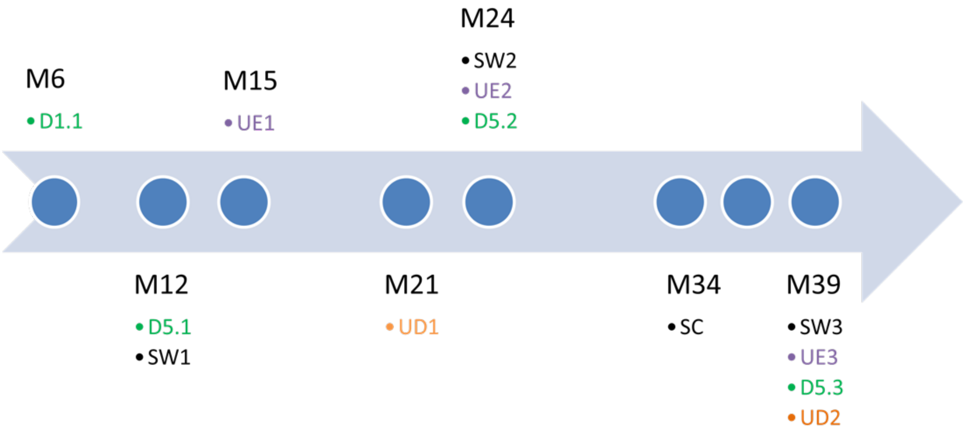


***Figure 75** Broad Timeline of Evaluation Activities Planned*

Timeline Legend:

- D = deliverable
- SC = scalability testing
- SW = software release
- UE = user evaluation (usability)
- UD = user day

Software releases and user evaluation ware done as planned. Our first SELMA User Day was held on 12 October 2022 in Bonn. The second (on-site) User Day was on 14-15 November 2023 in Avignon. Our final (virtual) User Day was on 21 March 2024.

6. Conclusion

This document follows up on D5.2 - Evaluation Plan, describes the evaluation activities in Y3 of the SELMA project and provides an overview of the evaluation over the entire project period. It combines contributions from all consortium partners and relates to other deliverables, including D1.1 - Use Case Descriptions and Requirements, D1.4 Final Prototype Report, D2.7 - Final progress report on continuous massive stream learning, D2.8 - Final release of continuous massive stream learning tools, D3.7 - Final report on speech and natural language processing, as well as D3.8 - Final release of speech and natural language processing tools, D4.4 - Final platform release with full continuous massive stream learning capabilities.

It provides an update of the list of components that were evaluated.

In this phase, technical evaluation focused on the final modules and platforms of all technologies developed in SELMA, including ASR enhancement, diarization and speaker recognition, speech-to-translated-text, speech synthesis, NER/NEL, summarization, integration of demonstrators and orchestration.

User assessment efforts targeted benchmarking of transcription, translation and voice-over engines, NER analysis, speech-to-translated-text, usability evaluation of the three primary demonstrators plain X, Monitio and the OSS, and the use case applications on speaker, podcasting and diversity.

This D5.3 Final Evaluation Report uses input from D5.1 - Evaluation Plan and D5.2 Interim Evaluation Report.

The evaluation shows that the language technology components developed within plain X are state of the art and reveals how fast such language technologies improve. The SELMA integrated platforms and Use Case Applications demonstrate their potential and show how they perform in a productive environment or at least have been trialed there. It also proves how industry as well as the academic world benefits from such research and innovation work.

Annex

Acronyms

Below is a list of acronyms that are used in this deliverable.

Acronym	Expansion
API	Application Programming Interface
ASR	Automated Speech Recognition
AST	Automatic Speech Translation
BBC	British Broadcasting Corporation
BLEU	BiLingual Evaluation Understudy (measurement for MT)
BM	Benchmarking
chrF	Character n-gram F-score (measurement for MT)
Dx	Deliverable x
DW	Deutsche Welle
EBU	European Broadcasting Union
FhG	Fraunhofer Gesellschaft
FTI	Fast Track Innovation
IMCS	Institute of Mathematics and Computer Science
KPI	Key Performance Indicator
LIA	Laboratoire Informatique d’Avignon
Mx	Month x
MSx	Milestone x

MT	Machine Translation
NEL	Named Entity Linking
NER	Named Entity Recognition
NLP	Natural Language Processing
NYT	New York Times
OSS	(SELMA) Open-Source Software
PRIB	Priberam
RAI	Radiotelevisione Italiana
RIA	Research and Innovation Action
SC	Scalability Testing
Sc	Scenario
SEL	SELMA
SELMA	Stream Learning for Multilingual Knowledge Transfer
SW	Software Release
SWR	Südwestrundfunk (German broadcaster)
TER	Translation Error Rate (measurement for MT)
ToC	Table of Contents
TRL	Technology Readiness Level
TTS	Text-to-Speech
UCx	Use Case x
UD	User Day
UE	User Evaluation

UI	User Interface
UX	User Experience
WCAG	Web Content Accessibility Guidelines
WER	Word Error Rate (measurement for ASR)