plain X – AI Supported Multilingual Video Workflow Platform

Carlos Amaral[P], Catarina Lagrifa[P], Mirko Lorenz[D], Peggy van der Kreeft[D], Tiago Veiga[P]
[P] Priberam, Portugal, [D]Deutsche Welle, Germany

Abstract

plain X[1] is a web-based software tool for multilingual adaptation of video, audio, and text content. The software is a 4-in-1 tool, combining several steps in the adaptation process, i.e., transcription, translation, subtitling, and voice-over, all automatically generated, but with a high level of editorial control. Users can choose which engines are used, depending on the languages and tasks. They can range from the big tech (e.g., Azure, Google, OpenAI) to smaller players (e.g., Opentrad for translation between Galician and Portuguese). As a result, plain X enables a smooth semi-automated production of subtitles or voice-over, much faster than with older, manual workflows. The software was developed out of EU research projects and has already been rolled out for professional use. Based on the European origin, the balance between technology and human expertise is strongly considered – plain X brings Artificial Intelligence (AI) into the multi-lingual media production process, while keeping the human in the loop.

1 Introduction

Originally, plain X was built for media broadcasters, although its use has been extended to other sectors in need of language adaptations as well. A key driver is the growing amount of content which needs language adaptation, based on user or market needs, for enhanced accessibility and/or to comply with regulation. Feature development was initially based on the needs of Deutsche Welle (DW), a world broadcaster producing news content in over 30 languages[2]. The plain X platform is the result of a partnership between DW as user partner and Priberam, a Portuguese scaleup that develops AI powered products based in language technologies.

The platform simplifies the multilingual adaptation process to a large degree, enabling easy subtitling in source and any target language requirement. To ensure the best possible results the software strongly relies on a "human in the loop" approach, by providing editorial tools to combine AI and human language expertise. After being rolled out for daily use in Deutsche Welle, several other clients are already using plain X, based on a software-as-a-service (SaaS) subscription model.

2 Challenges

The concept for plain X originated from the need to produce more with less, i.e., to use automation in the production process, so media producers can increase the volume of certain target languages, distribute content in more languages, or use synthetic voice, allowing to reach more people in their own spoken tongue, including in specific African or Asian regions. Since the rollout, conversations with users show a strong and growing demand to better serve regions with more than one language (for example, Spain) or large groups of immigrants (Germany). So far, the sheer volume of work was often considered simply too high, which could change with workflow tools like plain X.

Another key element of plain X is that the platform is engine agnostic, foreseeing access to the best available engines now and in the future. As an example, DW produces content in so many

[1] https://www.plainx.com

[2] https://corporate.dw.com/en/multimedia-content-in-30-languages/a-15703976

languages, it is essential to cover as many languages as possible, in the best possible quality, through a combination of engines from carefully selected providers, for instance for transcription or translation. In plain X, users can freely switch between different transcription, translation and voice-over engines. The platform architecture allows for the update or inclusion of additional engines in a short time. The same happens with new features like diarization or voice cloning text-to-speech models. Automated benchmarking of the different models allows choosing the best default engine for each language.

Being engine agnostic by design has been key for the support of low-resource languages by integrating engines from smaller players like Opentrad, a translation engine between Spanish, Portuguese, Galician and Catalan or the one from Lesan, for major Ethiopian and Eritrean languages.

Tables 1 to 3 show the current engines and number of languages supported for transcription, translation and voice-over.

Engine	# Languages
Amberscript	39
Azure	77
Google	72
Selma	6
Speechmatics	50
Whisper	100
All engines	108

Table 1 - Current transcription engines

Engine	# Languages
Azure	111
Google	127
DeepL	30
Lesan	3
Meta (OSS)	100
Opentrad	4
UTran	3
All engines	165

Table 2 - Current translation engines

The number of supported languages in Table 2 are target languages. Considering all combinations of source and target languages, plain X currently handles more than 38.000 language pairs.

Engine	# Languages
Azure	74
Eleven Labs	29
Google	44
Selma	2
All engines	77

Table 3 - Current voice-over engines

plain X provides system default engines per language (pair)/task combination. These can be overridden per task, user or organization allowing full control of the models used. The system defaults are chosen according to evaluations of the available engines for a specific language or language pair. Whenever possible, these are made by native speakers. The plain X team proactively tries to get feedback from users of languages not yet used by anyone else, guiding them through the testing of the available alternatives. Whenever new models, new versions of existing models or new languages are added, the plain X team evaluates them and suggests the users of those languages to do the same. The system defaults are updated whenever needed to ensure the best quality at each moment in time for all users that just want to rely on the simplest yet powerful plain X experience.

As expected, most of the language quality issues were reported for low-resource languages such as Burmese, Somali and Tibetan. This has been a major driver for the plain X team to look for alternatives to “big” providers. That’s how translation engines like Lesan, Opentrad and UTran were spotted and integrated in plain X. These are just the first three because there are several small companies training such models, independently from the big technology providers. This is why we expect measurable progress for low-resource languages in the next 12 to 18 months.

As new clients started to use the platform, new features were added making plain X capable of dealing with subtitle templates for social media (portrait and square videos) and compliant with subtitling guidelines, for instance, from streaming platforms, through a flexible rule-based subtitle segmenter.

Integration with internal content management and publishing systems was also a key requirement to reach the highest level of user acceptance, beginning with DW.

3 Origin

plain X initially came out of the SUMMA multi-lingual media platform, funded by the European Commission's H2020 project as a basic prototype for controlled transcription and translation for media monitoring purposes within DW and BBC Monitoring.

This prototype was then further developed and funded through the Google Digital News Initiative projects speech.media and news.bridge.

Finally, Deutsche Welle, Germany's international broadcaster in need of such platform, and Priberam decided to turn the prototype into a scalable, fully operational multilingual platform for wider use, supporting the needs of broadcasters and other multilingual content producers. That was the birth of plain X, a platform which turns content from and into virtually any language.

From prototype to product launch, besides the development and improvement of new and existing features, the UI and scalability, legal requirements related to privacy and GDPR were also implemented in plain X, making it ready to the market.

4 Workflow

The goal-oriented workflow is easy to use, but very powerful, offering editorial users the comfort of their familiar workflow, yet encompassing advanced automated technologies to support them in the creative process.

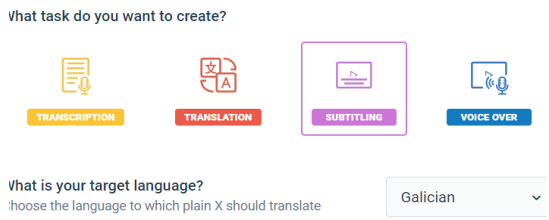

Figure 1 - Selection of goal for uploaded content

The first step is ingestion of content, be it video, audio, or text, with a growing number of supported input formats.

The next step for audiovisual content is transcription, through speech-to-text in the source language. That could be an end goal, for instance for interviews.

This also allows for a primary output of automatically generated source-language subtitles, which can be used as open or closed captions.

The next step is automated translation to a selected target language, which can be post-edited, something that is transversal to the entire platform – the human in the loop, one of the Responsible AI principles[3], that is, the user has always full control of the results. Again, the translation can be an end goal on its own, and used as input text for re-speaking, for example. One transcription can be translated to multiple languages.

However, it can also generate automated subtitling in the same target language. The subtitles are generated automatically taking into account not only the times given by the transcription engine but also a set of rules that can be customized. Again, the generated subtitles can be easily edited with the live preview of the subtitles in the video according to the selected template, for instance to avoid overlap of subtitles with name labels on the screen.

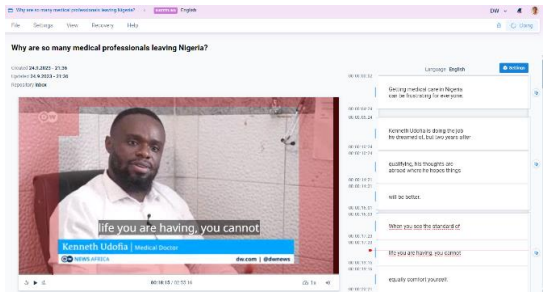

Figure 2 - Subtitling using templates

As a final step, the translation can be used for voice-over, by converting text to speech in the target language after selecting a synthetic voice. The user can also control the voice to a very high degree by applying simple commands to control the pronunciation, intonation and prosody of single words or full sentences.

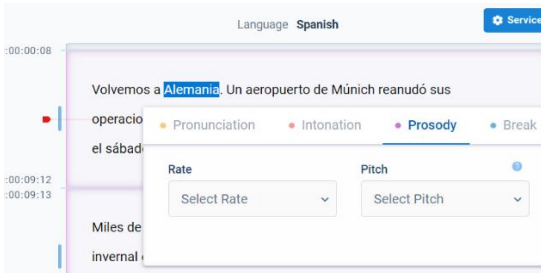

Figure 3 - Voice-over control options

[3] Priberam is a member of the Centre for Responsible AI (https://centerforresponsible.ai/)

Transcription and translation are the two most frequent tasks because one of the most used workflows in plain X is the transcription of audio interviews, followed or not by its translation. For video content, the combination of the sequence transcription -> translation -> subtitling is the one driving the rising usage of plain X answering a general media trend to have a much higher percentage of videos available with subtitling in a variety of languages.

Voice-over usage is still much lower because of the number of languages currently available (see Table 3) and the low quality of some of them.

The user reviewing and editing features embedded in all tasks are part of a human-centric platform that includes collaborative tools and workflows in every step, as required. Kanban-style boards clearly show the status of each task and the requests for input from a user, for instance, reviewing a translation.

Subsequently, other target languages can be added and produce equivalent content.

Estimates based on user surveys indicate time savings of 50% for transcriptions (speech to text), 30% for translations (here the human in the loop quality is the most important to reach high quality) and 70% for subtitling. These will surely rise with the quality improvement of the models.

Subtitling might be the most important task when combined with a more streamlined workflow and integration as plain X is simply faster compared to several tools.

5 Integration

It was vital to integrate this tool into the existing workflow infrastructure at Deutsche Welle and to allow for customization. This meant connecting it to input platforms for a smooth ingestion, as well as output tools for an efficient post-production and publication in the company style and branding.

Subtitle templates help to prepare the output in a particular house format. Other customizations include library management and access, setting subtitling rules, assigning roles to users, keeping track of usage and billing.

Working directly in a user environment from the start, with user input and feedback at every stage, allowed us to build a user-oriented platform to support editors in their adaptation process with the help of AI, while minimizing the feeling of insecurity and threat coming from automated processing.

The APIs used to integrate plain X with other systems can also be used for automation purposes like the fully automatic subtitling of video content from DW that is then pushed to Frankfurt Airport screens or the transcription and translation of video and audio content for media monitoring purposes. The voice-over engines in plain X can also be used to automatically generate podcasts from text content.

6 Future work

To cater for new use cases coming from current or new clients, a roadmap for the development of plain X is defined and is constantly being updated.

The “human in the loop” approach will likely gain relevance, because even at the now much higher quality of AI output, language must be treated with care and expertise.

Currently an internal benchmarking tool is used to compare the output quality of the different engines, which includes automated as well as user evaluation, and set the best rated engine as the default. We are currently working on a system where output labelled as final can also be used for benchmarking as well as training of engines and modules. As a result of the human in the loop approach and the growing number of users, the outputs from plain X will provide an increasingly reliable estimation of their quality. We will also test providing a tip to users suggesting to rerun a certain task with a different model whenever a certain threshold of mistakes is reached when editing the output of a transcription or translation task.

Some of the planned improvements already in development are the integration of quality estimation models, the ingestion of existing client terminologies and translations memories, speech Named Entity correction and the usage of LLMs to rewrite long subtitles, simplify text for more intelligible voiceover or customize the writing style of translations.

An overall policy is adherence to European privacy, data-protection and AI approaches as well as similar regulations in other geographies.

Acknowledgments

This work has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 957017, Project SELMA (https://selma-project.eu).