



X Modal
X Cultural
X Lingual
X Domain
X Site
Global OER Network

Grant Agreement Number:	761758
Project Acronym:	X5GON
Project title:	Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site Global OER Network
Project Date:	2017-09-01 to 2020-12-31
Project Duration:	40 months
Deliverable Title:	D5.3 – Final report on piloting
Lead beneficiary:	UPV
Type:	Report
Dissemination level:	Public
Due Date (in months):	40 (December 2020)
Date:	
Status (Draft/Final):	Draft
Contact persons:	Álex Pérez, Javier Jorge and Alfons Juan

Revision

Date	Lead author(s)	Comments
8-Dec-2020	Álex Pérez, Javier Jorge and Alfons Juan	first draft

Contents

1	Introduction	5
2	VideoLectures.Net	6
2.1	X5oerfeed component	6
2.2	X5recommend component	7
3	poliMedia	9
3.1	X5oerfeed component	9
3.2	X5recommend component	10
4	virtUOS	14
4.1	X5oerfeed component	14
4.2	X5recommend component	14
4.2.1	Goal and description	14
4.2.2	Method	14
4.2.3	Results	16
4.2.4	Discussion	20
4.2.5	Conclusion	22
5	Other pilots	22
5.1	Kobi app: helping children learn to read	22
6	User studies	23
6.1	Introduction to X5Learn	23
6.2	Overview of the X5Learn User Interface	23
6.3	Iterative Design and User testing	24
6.4	Content Flow Bar Pilot Study	24
6.4.1	Methodology	25
6.4.2	Procedure for the study	26
6.4.3	Results	26
6.5	Playlist Pilot Study	28
6.5.1	Introduction to Playlist	28
6.5.2	Playlist functionality	29
6.5.3	Designing X5Learn Playlists Tool: Initial user study	30
6.5.4	Pilot study	33
6.5.5	Results	35
6.5.6	Discussion	37
6.6	Conclusion: X5Learn	37
7	Advanced cross-lingual and cross-modal features	38
7.1	Streaming ASR	38
7.2	Simultaneous MT	39
7.3	Multilingual MT	41
7.4	Cross-lingual text-to-speech dubbing	43
7.4.1	The DeX-TTS dataset	44
7.4.2	Cross-lingual voice cloning at the UPV	45
7.4.3	Evaluation	47
7.4.4	Concluding remarks	51



8	Conclusions	52
A	virtUOS: additional details	57
A.1	Sample JSON structures	57
A.2	Test set structure	57
A.3	Recommendation Engine state and language structure	59
B	MediaUPV with multilingual subtitles as of June 2020	62



List of Figures

1	WER scores over time for VL.NET transcriptions in Sl and En	7
2	The distribution of position click in the X5recommender plugin.	8
3	Navigation between the providers through selected item amongst the recommendations.	8
4	WER scores over project month for poliMedia transcriptions in Spanish and English.	10
5	X5gon recommendations in poliMedia	11
6	Cross-site hits from poliMedia (UB="University of Bologna").	12
7	Cross-lingual hits from poliMedia.	13
8	Cross-modal hits from poliMedia.	13
9	UI draft of X5gon Discovery pilot (Screenshot)	15
10	Bar plot of crosstab for "Results Fit To Lecture" and "How Confident Are You"	18
11	Bar plot of "ResultsFitToLecture" for different model types.	19
12	Overview and clustering of user comments.	21
13	Hovering over a video fragment within the Content FlowBar	24
14	Swimlane view from video thumbnails	24
15	Usability (SUS) results	27
16	Playlist options after publishing	30
17	Playlist Creation, Annotating videos & adding video playlist	30
18	Search result for "Artificial Intelligence" in X5Learn initial interface	31
19	Automatically adding thumbnails to learning resources	32
20	Editing title and the description of an OER	33
21	Edit title and description view	33
22	Production pipeline of transcribed, translated and dubbed poliMedias.	46
23	Basic Tacotron2-UPV architecture.	46
24	Home page of the evaluation platform.	48
25	Naturalness evaluation interface.	49
26	Overview test set data.	58
27	Bar chart of content quantities for indexed language (top 7).	61
28	Diagram of the language structure of indexed content/OER.	61
29	A poliMedia studio (left) and example (right).	62



List of Tables

1	Frequency of navigation from VL.NET to other OER repositories	9
2	Recommendation hits comparison: UPV vs X5gon recommender.	10
3	Frequencies of languages per result record	17
4	Frequencies of duplicate or repetitive result records.	17
5	Statistics of Result Fit To Course/Lecture variable.	17
6	Frequencies of Result Fit To Course/Lecture variable.	18
7	Crosstab of variables "Results Fit To Lecture/Course" and "How Confident Are You".	18
8	Frequencies of "ResultsFitToLecture" for different model types.	19
9	Usability issues and solutions.	36
10	WER scores provided by offline and streaming-adapted M40 ASR systems.	39
11	Comparison between off-line baseline and simultaneous MT systems for Es→En	40
12	Comparison between off-line baseline and simultaneous MT systems for En→Es	40
13	Basic statistics (in millions) of the Slovenian-pivot multilingual training corpora . . .	42
14	Basic statistics (in millions) of the English-source multilingual training corpora. . . .	42
15	BLEU scores achieved by bilingual and multilingual Slovenian-pivot systems.	42
16	BLEU scores achieved by bilingual and multilingual English-source systems.	43
17	Statistics of data collection participants	44
18	Number of sentences and duration in hours of the clean speech data collected	45
19	Naturalness MOS with 95% confidence intervals per language	49
20	Speaker similarity MOS with 95% confidence intervals per language	50
21	Participant accuracy on the <i>real or synthetic</i> test.	50
22	Final questions and answers on the acceptance of TTS technology.	51
23	Test set data.	57
24	Result structure regarding model types and overlap.	58
25	Quantities and percentages of course languages.	59
26	Quantities and percentages of the test set faculties structure.	59
27	Structure of indexed OER per language for pilot 1, pilot 2 and their differences.	60
28	Number of poliMedia videos and hours in Spanish, Catalan and English.	63
29	Language statistics of poliMedia lecturers	63
30	WER/BLEU scores provided by UPV and Google ASR/MT systems on poliMedia . .	64

Abstract

The main objective of WP5 is to pilot successive versions of project components developed in WP1–WP4 by means of Task 5.1, *Piloting on individual components* (M6–M24, Leader UPV), and Task 5.2, *Piloting on integrated components* (M13–M36, Leader UCL). The work carried out in these tasks until M24 (August 2019) was reported in D5.1 (M6–M12) and D5.2 (M13–M24). D5.3 is to report the work done in Task 5.2 from M25 to M36 (Y3), and also during X5gon’s four-month extension to complete the user studies led by UCL (M37–M40). For simplicity, in what follows Y3 refers to the actual Y3 (M25–M36) plus the four-month extension (M37–M40), that is, from September 2019 (M25) to December 2020 (M40). As in D5.1 and D5.2, D5.3 first provides a brief summary of past work and an update on recent (Y3) work at each of the three official pilots individually (VideoLectures.Net, poliMedia and virtUOS). Other pilots are then covered in a similar way. This is followed by a detailed account of the work done in Y3 on the two main planned subtasks of Task 5.2. The first of these subtasks, led by UCL, is to pilot advanced analytics and social context meetings. To this end, UCL has conducted a number of user studies using the *X5Learn* platform also developed at UCL. The second of these subtasks, led by UPV, is to pilot advanced cross-lingual and cross-modal features. In this regard, UPV has explored the use of different techniques on the leading edge of knowledge in AI for natural language processing. In particular, excellent results are reported on OER for *streaming* automatic speech recognition, *simultaneous* and *multilingual* machine translation, and *cross-lingual* text-to-speech dubbing.

1 Introduction

Work Package 5 (WP5) was planned to carry out a series of piloting studies to link the automatic analysis undertaken on the data with the experiences of groups of learners. Generally speaking, these studies are expected to contribute significantly in revealing the factors that hold user engagement, make learning enjoyable and rewarding, and help in developing a rounded understanding of different disciplines.

The main objective of WP5 is to pilot successive versions of project components developed in WP1–WP4, namely:

- **X5oerfeed:** *The project will deploy a technological pipeline for content understanding that is based on wikifier, dmoz and other services developed by JSI, and video transcription and translation services developed by UPV.*
- **X5analytics:** *The project will track data of users and their progress and use that to drive an analytics engine driven by state-of-the-art machine learning that can improve recommendations through better understanding of users, their progress and goals, and hence their match with knowledge resources of all types.*
- **X5recommend:** *Cross-site and cross-lingual recommendation.*

The above does not include evaluating the use of the *X5gon platform* in the wild, or the coordination of a network of European OER repositories, which are main objectives for WP6 and WP7, respectively. Instead, WP5 can be seen as a primary source of feedback for WP1–WP4, and also for the non-technical WP6 and WP7.

WP5 runs from March 2018 (M6) to the (extended) end of the project (M40, December 2020), and consists of two main tasks:

1. **Task 5.1 Piloting on individual components (M6–M24, Leader UPV).**
Small in-house groups will be established to assess successive versions of project components developed in work packages 1 to 4. JSI will pilot individual components from WP2 and WP4, whereas UPV and Nantes will focus on WP3 components.



2. Task 5.2 Piloting on integrated components (M13-M40, Leader UCL).

UCL, JSI, UPV, UOS and Nantes will pilot integrated components in the social network. In M13-M24, they will start piloting advanced analytics and social context meetings both virtual and physical. In M25-M36, advanced cross-lingual and cross-modal features will be piloted for the social network to be prepared for different cultures.

This deliverable, *D5.3 – Final report on piloting*, is to report the work done in Task 5.2 from M25 to M36 (Y3), and also during X5gon’s four-month extension to complete the user studies led by UCL (M37–M40). For simplicity, in what follows Y3 refers to the actual Y3 (M25–M36) plus the four-month extension (M37–M40), that is, from September 2019 (M25) to December 2020 (M40).

At this point, it is worth to mention that, in M24, JSI finalised the integration of the three major (planned) project components (X5oerfeed, X5analytics and X5recommend) into the *X5gon platform* under different names. The reader is referred to deliverable *D2.2: Final Server Side Platform* for a detailed description of the X5gon platform architecture, ingesting and processing pipeline, database, services, API, and (X5gon) connect service [1]. Concerning WP5 and for coherence with D5.1 and D5.2, we still use the term X5oerfeed to refer to OER pipeline processing services, particularly (video) automatic transcription and translation services developed by UPV. Similarly, X5recommend refers to the X5gon recommender engine [1, Sec. 5.1]. Regarding the X5analytics component, whose development has been more difficult than anticipated, it is from M25 on (Y3) integrated into the X5gon platform through an API allowing access to multiple analytics, models and tools (see [1, Sec. 5.3], [2] and [3]). It is assessed, to some extent, as part of the user studies carried out by UCL.

The structure of this deliverable is as follows. As in D5.1 and D5.2, D5.3 first provides a final update of work done and results at each of the three official pilots individually: VideoLectures.Net in Section 2, poliMedia in Section 3, and virtUOS in Section 4. Other pilots are then covered in Section 5. This is followed by a detailed account of the work done in Y3 on the two main planned subtasks of Task 5.2. Section 6 covers the first of these subtasks, piloting advanced analytics and social context meetings, for which UCL has conducted a number of user studies using the *X5Learn* platform also developed at UCL. The second of these subtasks, led by UPV and covered in Section 7, is to pilot advanced cross-lingual and cross-modal features. In this regard, UPV has explored the use of different techniques on the leading edge of knowledge in AI for natural language processing (with excellent results on OER for *streaming* automatic speech recognition, *simultaneous* and *multilingual* machine translation, and *cross-lingual* text-to-speech dubbing). Finally, the main conclusions drawn in Y3, for Y3 and also the whole duration of WP5, are provided in Section 8.

2 VideoLectures.Net

2.1 X5oerfeed component

Being the largest official pilot in X5gon, Videolectures.NET has been a primary focus of interest to WP5 during the whole project. In particular, the quality of automatic transcriptions and translations for Videolectures.NET videos, most in English and Slovene, has been considered a major objective since the very beginning of X5gon. As described in [4, Section 2], in M24 we had already achieved good scores for transcription error in English (WER below 20%) and Slovene (WER of 25.3%). Nevertheless, as described in [5, Section 2.2], from M25 to M30 we reduced the transcription error in Slovene (from 25.3% to 22.0%). Moreover, thanks to late developments for advanced cross-lingual features (Section 7.1), we have further reduced transcription error rates, both in Slovene and English, down to an impressive WER of 15%. Summarizing, Figure 1 shows the WER scores achieved over project month for Videolectures.NET transcriptions in both languages.



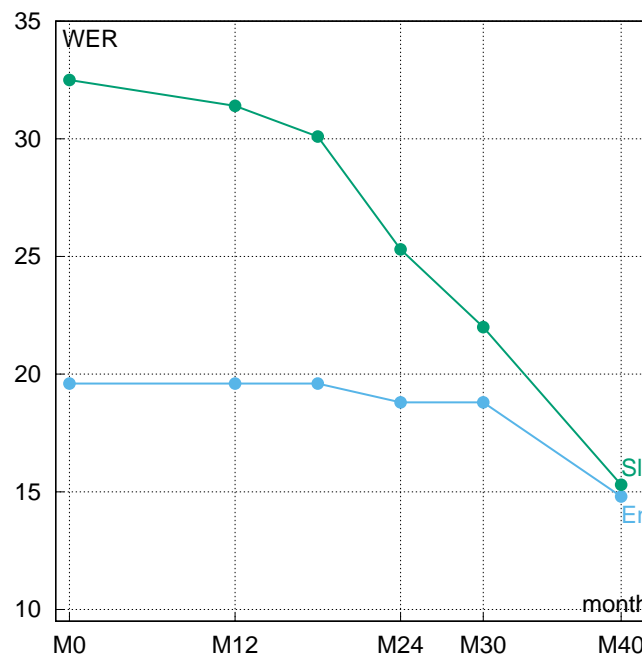


Figure 1: WER scores over project month for Videolectures.NET transcriptions in Slovene and English.

From the results in Figure 1, it can be concluded that the overall progress on transcription accuracy for Videolectures.NET videos in English and Slovene has been excellent. As discussed in [5, Sections 2.1 and 2.2], based on our prior research experience in ASR and MT for OER, WER scores around 20% or below are in general of publishable quality and, not least, suitable enough to try MT from them. Thus, crossing into the “safe” area of WER scores below 20% is clearly a major breakthrough for Videolectures.NET and X5gon. For the reader to get an idea of how difficult this AI challenge is, Google Cloud Speech-To-Text on Videolectures.NET videos only attains WER scores of 28.6% and 50.0% in, respectively, English and Slovene (see [5, Section 2.2]).

Regarding the quality of automatic translations for Videolectures.NET videos, in M24 we had already obtained very good results for translation accuracy between English and Slovene. Indeed, in the comparison to Google Translate reported in [5, Section 3.3], our results (in terms of BLEU scores) achieved relative improvements of 76% for Slovene→English and 39% for English→Slovene. From M25 to M30, we focused on improving X5gon MT systems for English↔{Spanish, French} on Videolectures.NET, with relative improvements around 10% over M24 scores, though more or less on par with Google Translate [5, Sections 3.2 and 3.3].

2.2 X5recommend component

Since the last time reporting, the database that stores the user transitions via the X5recommend plugin has increased for about 1 million records to a total of 1,281,040 records. These records show how a user navigated from one page to another through the recommended items. The recommended list usually contains around 20 items. Through the analysis of these records we found that the users tend to choose the item with an average rank 9.6. Similarly as before, only 5 to 7 items could fit into the recommendation window, that is why the users still scroll down in the window and choose one of the items.

According to the statistics, the users have chosen an item from the first page 448,079 (34.98%) and have scrolled down to choose an item 832,960 (65.02%) in the scenario where 6 items are shown

at once in the recommendation window. Figure 2 show the overall distribution of clicks per item rank in the X5recommender plugin.

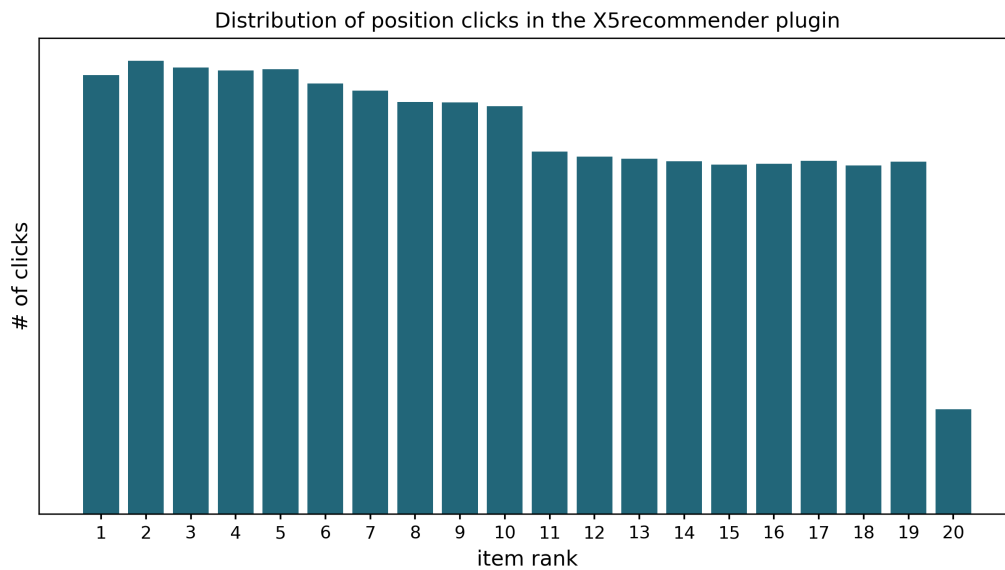


Figure 2: The distribution of position click in the X5recommender plugin.

Since the recommendations are cross-domain, it is possible for the users to move from one OER repository to another. Because of the data sharing policy among the providers, we can track the directions from VideoLectures.NET to any OER provider. The Sankey diagram in Figure 3 shows the navigation among the OER repositories.

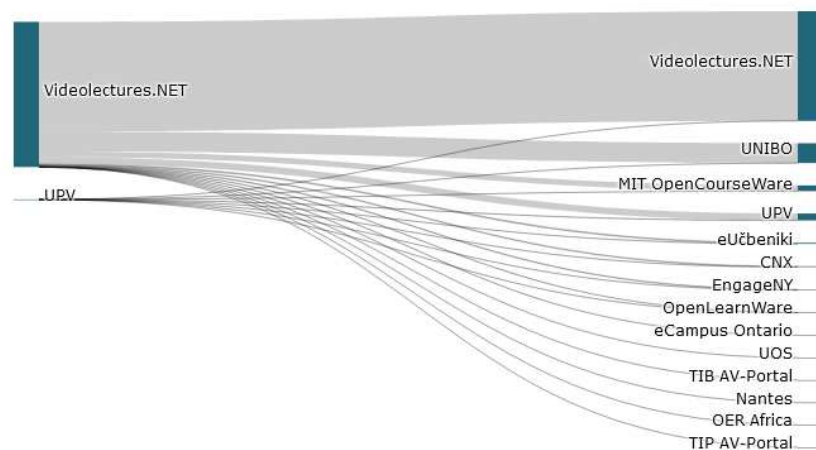


Figure 3: Navigation between the providers through selected item amongst the recommendations.

Table 1 shows the exact number how many times a user is directed from one repository to another. These results can be concluded as:

- When a user is on a particular material, most probably they choose the next material from the same domain so they prefer to stay where they are already.

- The users have mostly chosen the next item from the Videlectures.NET, UNIBO (Bologna), UPV and MIT respectively.
- All transactions to Nantes, UOS, MIT, OER Africa, TIB AV-Portal were directed from Videlectures.NET.

Directed From	Directed To	Abs Count	Ratio
Videlectures.NET	Videlectures.NET	917,633	75.541% (75.449%)
Videlectures.NET	UNIBO	165,918	13.659% (13.642%)
Videlectures.NET	UPV	57,222	4.711% (4.704%)
Videlectures.NET	MIT OpenCourseWare	47,616	3.920% (3.915%)
Videlectures.NET	eUčbeniki	8,757	0.721% (0.720%)
Videlectures.NET	CNX	6,576	0.540% (0.540%)
Videlectures.NET	EngageNY	5,241	0.431% (0.431%)
Videlectures.NET	OpenLearnWare	2,188	0.180% (0.180%)
Videlectures.NET	eCampus Ontorio	2,005	0.165% (0.165%)
Videlectures.NET	UOS	1,167	0.096% (0.096%)
Videlectures.NET	TIB AV-Portal	373	0.031% (0.031%)
Videlectures.NET	Nantes	42	0.003% (0.003%)
Videlectures.NET	OER Africa	12	0.001% (0.001%)

Table 1: Frequency of navigation from VL.NET to other OER repositories. The “Ratio” indicates the frequency of transitions that happened from the given “Directed From” repository, while the value in brackets indicate the frequency of transitions between the OER repositories across the whole data set.

3 poliMedia

3.1 X5oerfeed component

As with Videlectures.NET, the quality of automatic transcriptions and translations for poliMedia videos, most in Spanish but also in English, has been considered a major objective since the very beginning of X5gon. In M24, the WER scores for the automatic transcription of poliMedia videos were already well below the 20% “safety” threshold: 11.0% for Spanish and 15.8% for English. However, as we did for Slovene in Videlectures.NET, from M25 to M30 we reduced the transcription error in Spanish to 9.1% [5, Section 2.2]. Later on, also as we did for Videlectures.NET, we managed to further reduce transcription error rates for Spanish and English down to 8.3% and 12.0%, respectively, by taking advantage of late developments for advanced cross-lingual features (Section 7.1). Figure 4 shows the WER scores achieved over project month for poliMedia transcriptions in Spanish and English.

The excellent results in Figure 1 clearly show that high-quality automatic transcriptions of poliMedia videos are now a genuine reality. In M30, we tested Google Cloud Speech-To-Text on poliMedia videos in Spanish and got a WER score of 19.9%, that is, our 8.3% means a relative improvement of 58.3% [5, Section 2.3]. It goes without saying that this is major achievement for X5gon in terms of AI technology.

On the translation side, our M24 BLEU score of 30.0 for the translation of poliMedia videos from Spanish to English was improved up to 34.1 in M30 [5, Section 3.3]. At first sight, it might seem that this improvement, 13.7% relative, simply allows us to approach the 35 threshold commonly used by experts to consider automatic translations good enough for practical use. However, it is actually a

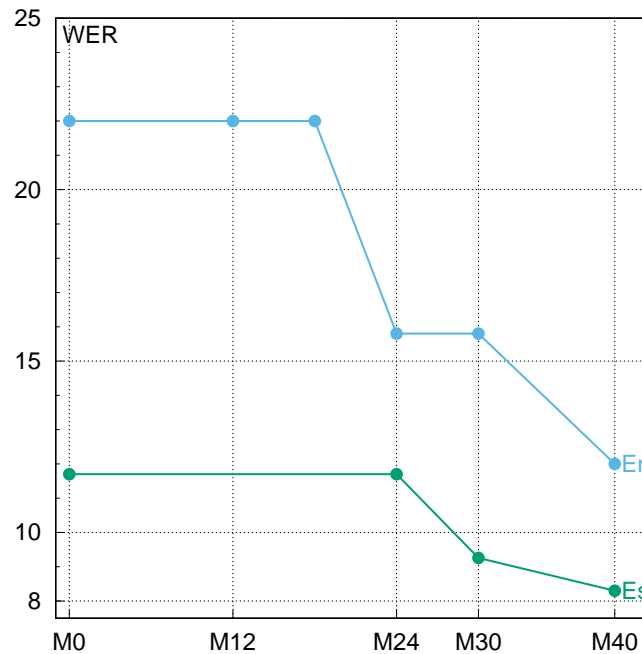


Figure 4: WER scores over project month for poliMedia transcriptions in Spanish and English.

major step forward since, combined with very low 8.3% transcription error for Spanish, this level of translation accuracy means that, for the first time, we truly have a fully automatic pipeline to reliably translate poliMedia videos from Spanish to English.

3.2 X5recommend component

Since the first half of Y2, the X5recommend component has been providing cross-lingual, cross-modal and cross-site recommendations to poliMedia students (see Figure 5) with the main objective of collecting and analyzing their interactions with the X5gon recommendations, and also to use this valuable user data to adapt the models and tools developed in X5gon. The collected data has been used to further improve the recommendations, to train and adapt the learning analytics models, and also to assess how poliMedia students can benefit from having related OER materials available.

The X5gon recommendations and the poliMedia official recommendations are randomly shown in a 50-50 ratio to students accessing the UPV media portal. The total number of user clicks logged for each of the recommenders is 1454 for the UPV's, 712 for the X5gon's. These numbers correspond to logs from M19 (March 2019) to M35 (July 2020).

To begin with, a simple comparison on the usage of both recommenders (the official poliMedia recommender versus the X5gon recommender) in terms of user clicks is given in Table 2.

Recommender	Hits	Percentage
UPV	1454	66.7%
X5gon	712	33.3%

Table 2: Recommendation hits comparison: UPV vs X5gon recommender.

From the figures in Table 2, it can be seen that poliMedia students are slightly more attracted by the UPV official recommendations, which only contains other poliMedia OER as possible recommen-

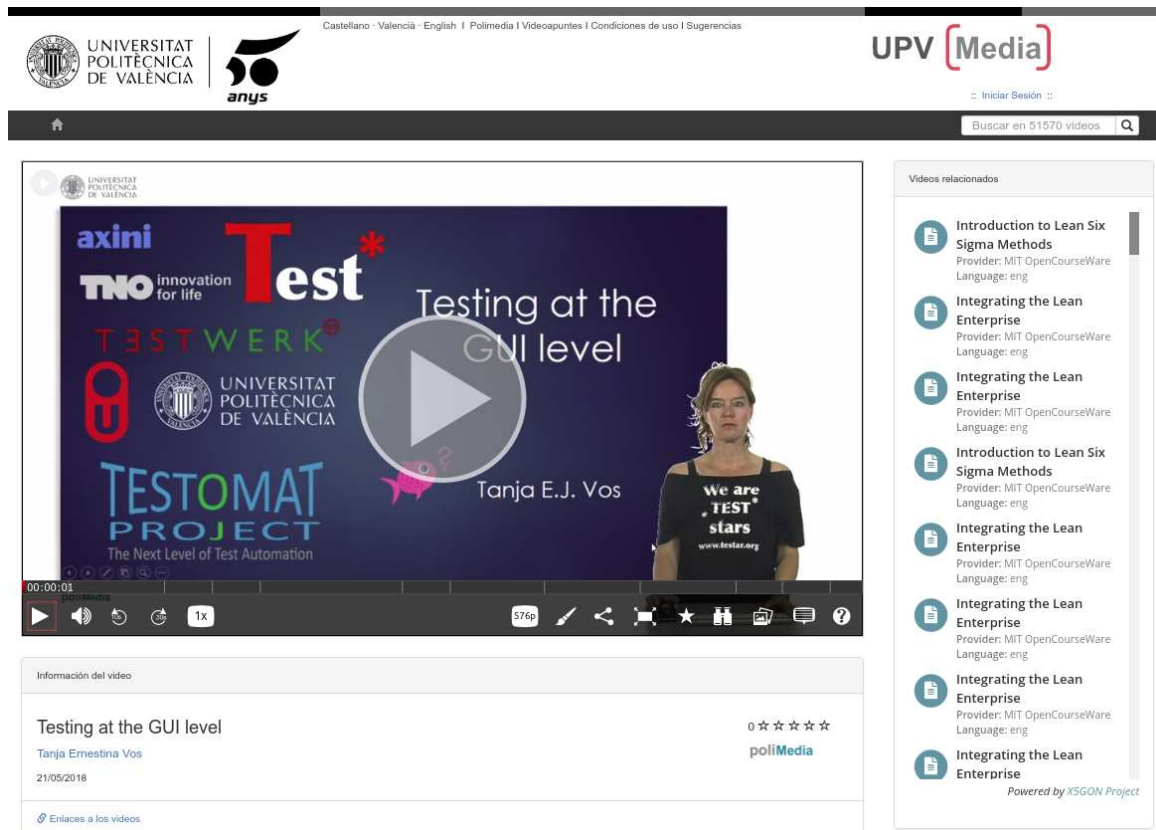


Figure 5: X5gon recommendations in poliMedia

dations. This can be explained by the way UPV students consume OER contents in the UPV media portal. In general, UPV students use poliMedia objects to learn about very specific topics (usually provided in lecture pills of no more than 10 minutes), specially when preparing for an examination, and thus they might find other poliMedia OER resources more suitable for their needs. We should also not forget the fact that the variety of OER resources available in the X5gon network is yet rather small, and specific topics on different areas might not be covered at all. Nevertheless, there is a positive acceptance and engagement from UPV students on having related OER materials recommended by X5gon.

Next we provide an analysis of cross-lingual, cross-site and cross-modal OER recommendation links followed by UPV students on the poliMedia site. Providing students with OER recommendations in a variety of languages, sites or formats is one of the main motivations of X5gon, and thus it is important to pay attention to how UPV students make use of such contents when provided by the X5gon recommender in a real-life learning scenario. When analyzing these numbers, we should keep in mind the aforementioned particularities on OER consumption presented by UPV students.

First, we would like to analyze cross-site events (poliMedia students following a recommended OER from a different site), as they show evidence of the described behaviour of UPV students. Figure 6 shows that more than 90% of the recommendations followed by poliMedia students point to other poliMedia OER contents.

Another key aspect of X5gon regarding OER contents is the language and the language barrier. Figure 7 shows the distribution of poliMedia students following recommendation links of different languages. As it can be seen, more than 80% of the recommendations followed by students are from Spanish/Catalan OER to Spanish/Catalan OER. The majority of the cross-lingual events are from

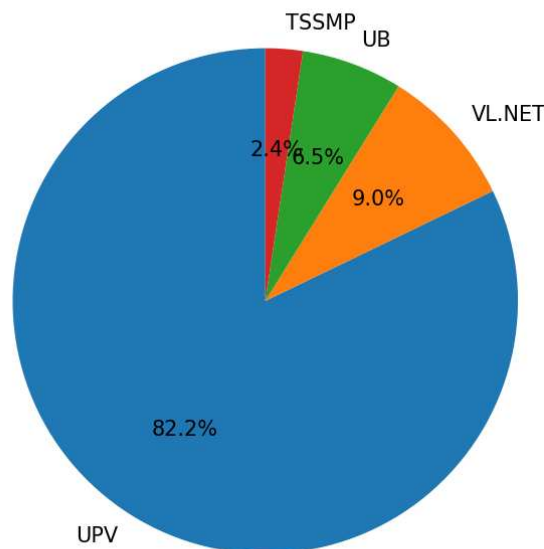


Figure 6: Cross-site hits from poliMedia (UB="University of Bologna").

Spanish/Catalan to English. This is again significantly influenced by the fact that UPV students find other poliMedia OER contents more suitable for their needs, and the vast majority of them are provided either in Spanish or Catalan. Similarly, the preference for English on cross-lingual events can be influenced by the wider availability of contents in that language, and not only explained by the language factor. Under this scenario, it is difficult to extract solid conclusions on the role that the language in which the OER contents are given can be playing.

Although the UPV portal only provides learning materials as video lectures, OER contents exist also in different formats (documents, web pages, etc). The X5gon recommendation frame that poliMedia students are shown presents the recommendation links accompanied by different icons depending on the document format of the suggested OER, so the student knows beforehand the type of the suggested material. We find also interesting to pay attention to the recommendation events that commute between formats (for example, accessing a related PDF document while following an OER video lecture). Figure 8 shows that cross-modal events are about 10% of the total, specifically from a poliMedia learning pill to an external PDF document.

Given all these figures, we hypothesize that the fact that the suggested OER belongs to the same UPV course/subject is the main factor that influences UPV students when following a recommendation link. The scarce variety of topics currently covered by the X5gon network and the particular way UPV students consume poliMedia video lectures to prepare exams make it difficult to draw solid conclusions on having cross-site, cross-lingual and cross-modal OER related contents available.

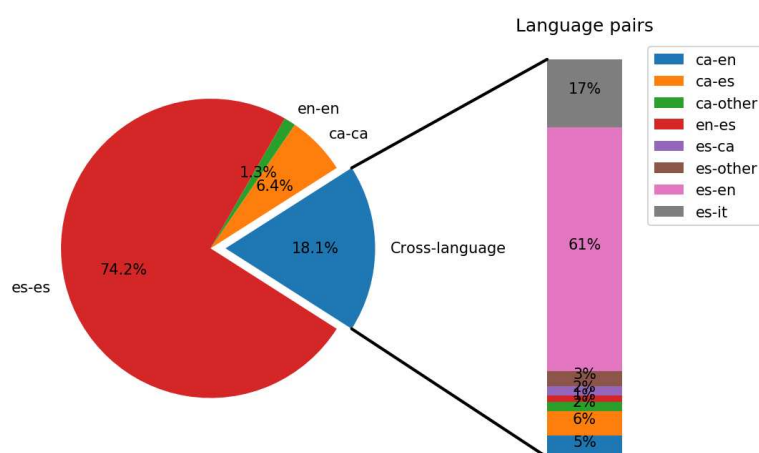


Figure 7: Cross-lingual hits from poliMedia.

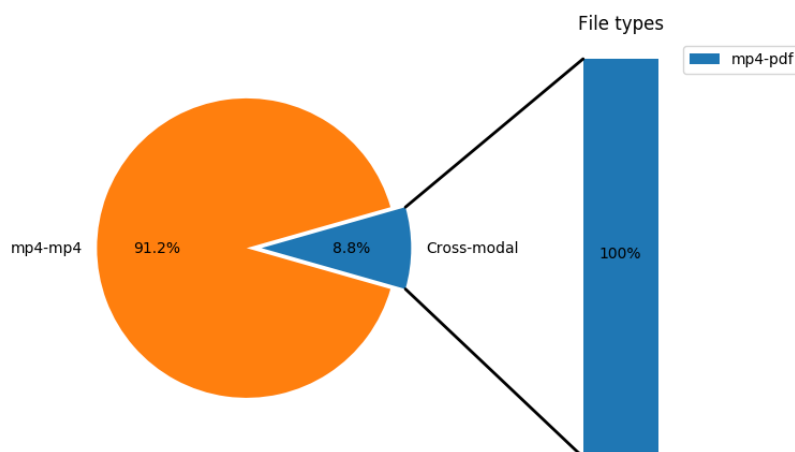


Figure 8: Cross-modal hits from poliMedia.

4 virtUOS

4.1 X5oerfeed component

During Y1 (M6-M12) and the beginning of Y2, UOS finalized the integration of the (X5gon) connect service into its virtUOS institutional repository. All the OER contents from virtUOS were automatically transcribed and translated. Evaluation sets were defined in order to carry out the transcription and translation quality evaluations through the UPV's Transcription and Translation Platform. Evaluation results were reported in [6, Section 4] and [4, Section 4]. Given that UOS's main interest was to pilot OER recommendations to lecturers, no further developments and evaluations of the X5oerfeed component were planned for UOS in Y3.

4.2 X5recommend component

As discussed in [4, Section 4], in Y2 UOS updated its Y1 *OER Recommender for Lecturers* pilot proposal to the *X5gon Discovery Pilot* detailed in [4, Sections 4.2.2 to 4.2.6]. In Y3, UOS run a new edition of this pilot which is referred to as *X5gon Discovery Pilot 2*. As in [4, Sections 4.2.2 to 4.2.6], this new edition is described below in terms of goal and description, method, results and conclusion (with previous discussion).

4.2.1 Goal and description

X5gon OER recommendation and search engine is in a prototype state since 2018. To achieve long-term goals and to get potential users to use the search engine, it is necessary that the search engine provides valuable results. Search results have to be related to the search query and should also be relevant for the users search goal or learn goal. If users are not able to find OER related to their search topic or if it's too hard to get a useful result out of it, the users won't be willing to use the X5gon search engine in future.

X5gon Discovery should therefore be designed in such a way that users can quickly access relevant information or reach their desired search destination. As described in [7], a qualitative analysis of the search results is necessary for this purpose. Since the vast majority of the users view only the first 10 search results [8], these results need to be the focus of the evaluation. Since this has already been discussed in more detail in last year's X5gon Discovery Pilot documentation, it will not be addressed any further at this point.

X5gon Discovery Pilot 1 used data from the search engine of the Jozef Stefan Institute (JSI). In this pilot, data from the search engine developed in Nantes is used to evaluate the material results. Furthermore, this pilot asks the question about the search engine's result quality for different model types. At the time of the execution of the pilot three different model types were implemented in the Nantes search engine: *tfidf*, *wikifier* and *doc2vec*. This allows to choose a suitable model type in the long term and give the users the best possible search results.

4.2.2 Method

UOS used course data from Stud.IP (UOS Learn-Management-System) like title and description and feed the X5gon Discovery API with that data. For each search string (title + description) searches are performed for each model type, grouped by duplicate results and saved. Per search string (or course metadata) 10 to 15 results were provided. These results were saved and evaluated by members of the target group (lecturers or students in higher education context), who were asked the following:

- The search result matches the content of the lecture/course.

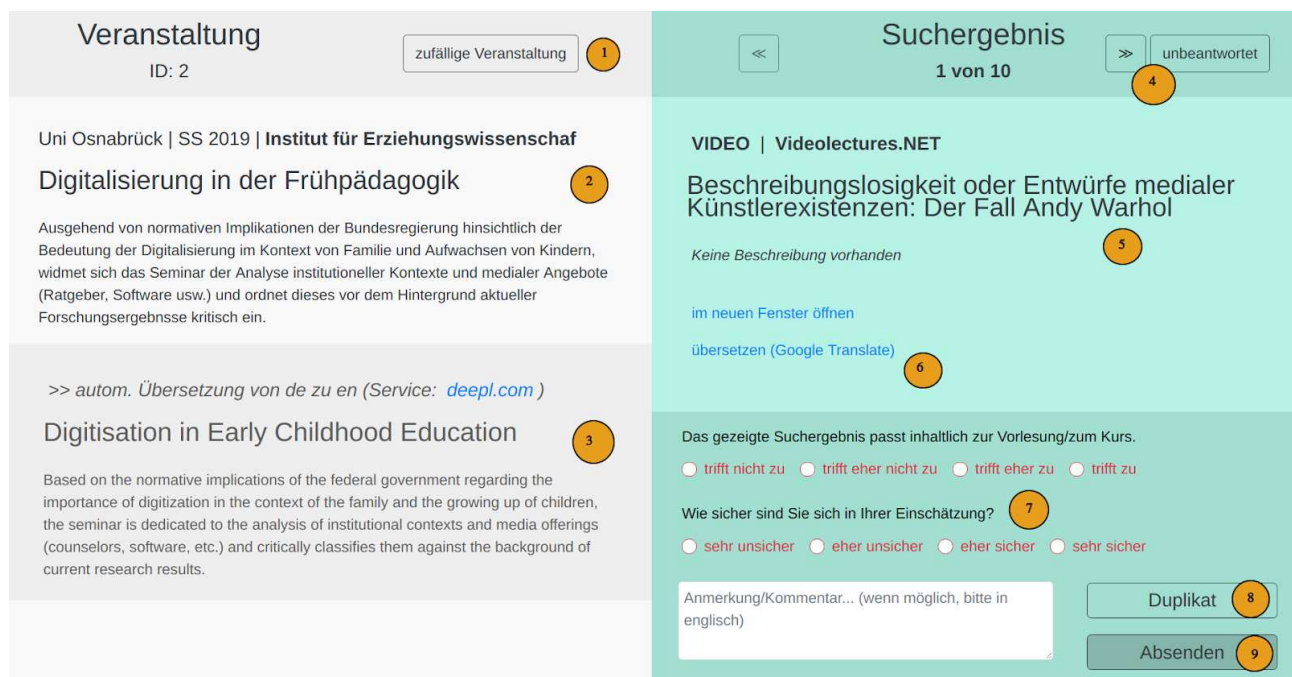


- How sure are you of your evaluation?
- (optional) Result is a duplicate (no evaluation needed).
- (optional) Comment.

Survey results on Likert scales were saved per lecture and per result in JSON format to ensure sufficient evaluation potential. For the evaluation process, a web application with a corresponding back-end was developed by UOS. This tool meets all requirements regarding the evaluation of search results, as it is easy-to-use for the target group and was also used for Discovery Pilot 1.

Web application and technical details

The Web UI is split roughly into two columns. On the left are data about the course such as title, description and institute. On the right side, stored search results for the selected course are displayed.



The screenshot shows a web application interface divided into two main sections. The left section, titled 'Veranstaltung' (Event), displays details for a course at Uni Osnabrück. It includes a 'zufällige Veranstaltung' (random event) button (1), the course title 'Digitalisierung in der Frühpädagogik' (2), a detailed description in German (3), and an automatic English translation (3). The right section, titled 'Suchergebnis' (Search result), shows a search result for 'Beschreibungslosigkeit oder Entwürfe medialer Künstlerexistenzen: Der Fall Andy Warhol' (5). It includes navigation buttons (4), a 'Keine Beschreibung vorhanden' (no description available) message (5), a link to 'im neuen Fenster öffnen' (open in new window) (6), and a Google Translate link (6). Below the search result, there are evaluation options (7) and a 'Duplikat' (duplicate) button (8). At the bottom, there is a comment field (9) and an 'Absenden' (send) button (9).

Figure 9: UI draft of X5gon Discovery pilot (Screenshot)

1. "Random event" button: automatically selects a random event mouse
2. Description of the selected course.
3. Translation of the course data into English to overcome language barriers such as French, Italian or Spanish.
4. Navigation of search results (the "unanswered" button jumps to the next search result not yet rated by you).
5. Description of the selected search result (the link "open in new window" leads directly to the resource).
6. "Translate" button: opens title and description in Google Translator.
7. Comment field: here is space for short comments and remarks, which help to evaluate the data. Any other symbols, formatting etc. in the title and the description of the search result should be

entered here. Comments refer to the currently selected search result. Example: "unintelligible symbols in description". Example: "html-tags in title".

8. "Duplicate" button: if a search result has already been listed, a duplicate can be reported here.
9. "Submit" button.

Sample JSON structures are provided in Section A.1 for stored lecture and search results data, and also for the surveyed data.

Test set structure

In order to ensure the comparability of the pilot results, the same course data from the 2019 summer semester were used for "Discovery Pilot 2" as for "Discovery Pilot 1". There are, however, minor deviations due to the new query approach and come from the combined search results (for all three model types). Section A.2 provides an updated detailed description of the test set structure.

Recommendation Engine state and language structure

With the aim of comparability in further evaluation phases in the future, the data of the Recommendation Engine was extracted at the time of the search result name. The total number of indexed OER at the time of search result name in calendar week 04 in 2020 is $n = 111975$. At the time of data retrieval for Discovery Pilot 1 (calendar week 19 in 2019), a total of $n = 88295$ OER were indexed in the database. This is an increase of 23680 indexed OER items in the X5gon database (+26.82%). A detailed analysis of the language structure in indexed OER is provided A.3, which of course is largely dominated that by that of the official pilots.

4.2.3 Results

This section contains an overview and analysis of data collected in the phase of the X5gon Discovery Pilot 2 at UOS from the 8th to the 22nd of January 2020. As discussed above, the main goal of the pilot and the analysis is the evaluation of X5gon search- and recommendation engine regarding quality and relevance of the top 10 search results through our target group (students and lecturers). Some relevant additional details are as follows:

- Date of survey data collected (X5gon API): 22nd of January 2020.
- Query parameters: title and description of 22 current UOS courses (text string).
- A total of 305 search results were evaluated (distributed over 22 courses).
- Three different model types were differentiated: "tfidf", "wikifier" and "doc2vec". Duplicate results were grouped.
- Data set has $n = 1617$ evaluations (~ 73.5 per course; ~ 5.3 per search result).

Result Language Structure (languageCode)

This variable shows the language structure of the evaluated search results. UOS mainly offers German and English courses, though some courses are also offered in other languages such as Spanish, French and Italian (see Section A.2 for details). The search result evaluation data, shown in Table 3, are therefore representative for the language structure of the UOS courses.

Compared to the language structure of the test set, given in Table , there are 4% fewer German results, which in turn is reflected in a slight increase in the other languages. Beyond that there are no notable differences between the language distribution in the test set (also see "Test set structure") and the percentages shown in Table 3.



Table 3: Frequencies of languages per result record

Levels	Counts	% of Total	Cumulative %
de	815	50.4	50.4
en	554	34.3	84.7
es	84	5.2	89.9
fr	82	5.1	94.9
it	82	5.1	100.0

Is the search result a duplicate? (isDuplicate)

In contrast to the first pilot, the focus in this pilot was on the three different model types described above. By previously grouping the search results, duplicate items (by materialId) were already sorted out. Nevertheless, 9% results were marked as duplicates by the users (Table 4).

Table 4: Frequencies of duplicate or repetitive result records.

Levels	Counts	% of Total	Cumulative %
false	1471	91.0	91.0
true	146	9.0	100.0

Search result matches the content of the lecture/course

All entries marked as duplicates (also see “Is the search result a duplicate? (isDuplicate)”) have been sorted out resulting (minus 146 survey results) in a sample count of $n = 1471$. Regarding the quality of the evaluation, one lecture ($id = 21$) was completely filtered out with 82 survey results, resulting in a total number of evaluable survey results of $n = 1397$. Participants were asked if the search results shown matched the content of the corresponding course and evaluated using a Likert scale with answer options: (1) “Strongly disagree”, (2) “Disagree”, (3) “Agree” and (4) “Strongly agree”. Table 5 shows basic statistics of the results. It is worth noting that the mean of the evaluation is 2.06, with a standard deviation of ± 1.03 , and thus lying between (1) “Strongly disagree” and (3) “Agree” with a tendency to (2) “Disagree”.

Table 5: Statistics of Result Fit To Course/Lecture variable.

N	Mean	Median	Std. Dev.
1397	2.06	2	1.03

Table 6 shows the counts and proportions for each possible opinion. On the one side, 63.9% of the search results were rated by participants as not matching the corresponding course (Disagree + Strongly disagree). On the other side, 36.1% of the search results were rated by participants as matching the course (Agree + Strongly agree). More than a third of the search results (39.8%) were also rated as absolutely inappropriate. In contrast, 9.9% of the search results were rated as very suitable.

Crosstab: ResultsFitToLecture / HowConfidentAreYou

All entries marked as duplicates have been sorted out resulting in a sample count of $n = 1397$. In order to evaluate the validity of the data with regard to the fit to the course, the participants were additionally asked how confident they were with their assessment. Table 7 shows a crosstab of the variables “ResultFitsToLecture” and “HowConfidentAreYou”. In Figure 10, it is shown in percentages.



Table 6: Frequencies of Result Fit To Course/Lecture variable.

Levels	Counts	% of Total	Cumulative %
(1) Strongly disagree	556	39.8	39.8
(2) Disagree	336	24.1	63.9
(3) Agree	366	26.2	90.1
(4) Strongly agree	139	9.9	100.0

Table 7: Crosstab of variables "Results Fit To Lecture/Course" and "How Confident Are You".

ResultFitsToLecture	HowConfidentAreYou			
	Very uncertain	Uncertain	Certain	Very certain
Strongly disagree	48	29	117	362
Disagree	66	73	144	53
Agree	34	122	158	52
Strongly agree	1	3	52	82

ResultsFitToLecture/HowConfidentAreYou in percentages

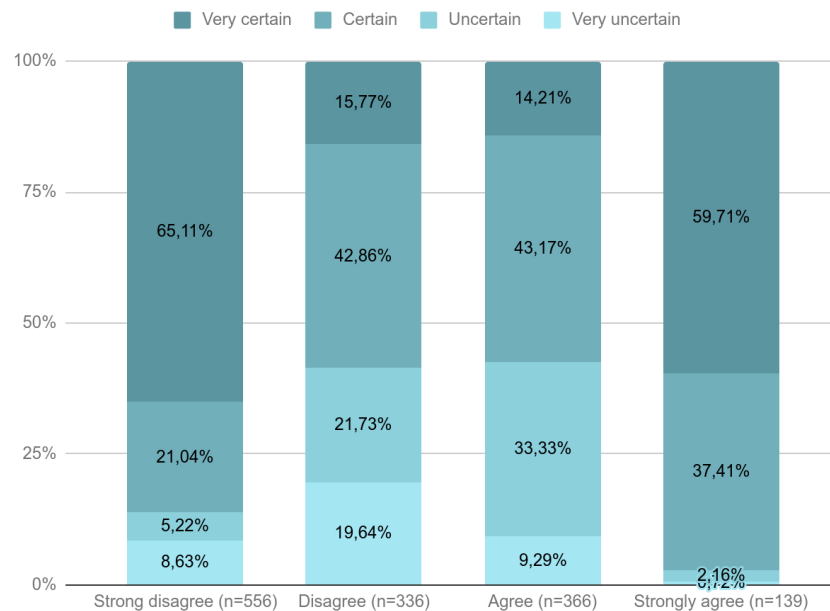


Figure 10: Bar plot of crosstab for "Results Fit To Lecture/Course" and "How Confident Are You".

From the plot in Figure 10, it can be seen that the frequencies of samples of the individual bars varies widely (see frequencies in the axis labeling). It can be recognised that the participants were more than 97% certain or very certain about their assessment when rating "Strongly disagree" (59,71% + 37,41%) and ~86% "Strongly agree" (65,11% + 21,04%). When rating "(3) Agree" or "(2) Disagree" in the middle range of the scale, users indicate that they are rather unsure about their ResultFitToLecture rating. For "(3) Agree", users rated their choice as 33.33% "Uncertain" and 9.29% "Very uncertain", which is 42.6% overall. A similar distribution can be seen in "(2) Disagree", where users indicated 21.73% "Uncertain" and 19.64% "Very uncertain" (total: 41.4%). The results for "Very certain" are 14.21% for "(3) Agree" and 15.77% for "(2) Disagree".

To understand the differences of the model types regarding the "ResultFitToLecture" quality, Table 8 provides a comparison of the variable "ResultsFitToLecture" with the corresponding occurrences for the model types. A total of $n = 1521$ ratings were given and for each of the three model types, an average of ~ 507 ratings were recorded. The reason for the deviation at this point from the filtered results ($n = 1397$) described above is the number of search results suggested by several model types (for example, by "tfidf" and "wikifier" on the same search string). This deviation of approximately 8.15% is similar to the 7.6% overlap of model types described above on the test set structure.

Table 8: Frequencies of "ResultsFitToLecture" for different model types.

ResultFitToLecture	tfidf	wikifier	doc2vec
(1) Strongly disagree	150	178	261
(2) Disagree	130	118	120
(3) Agree	155	141	105
(4) Strongly agree	70	61	32
Sums	505	498	518

Figure 11 shows a bar plot illustrating the differences among the different model types. For "tfidf" the sum of 44.6% results from the user ratings of 13.9% "(4) Strongly agree" and 30.7% "(3) Agree". Users rated the results of the wikifier model at 12.2% with "(4) Strongly agree" and 28.3% with "(3) Agree", giving a total of 40.5%. The results weighted by the doc2vec model received a total score of 26.5%, consisting of 6.2% "(4) Strongly agree" and 20.3% "(3) Agree". The difference between these cumulative ratings is 4.1% from the best-rated results of the model "tfidf" to second place "wikifier". The difference between "tfidf" and "doc2vec" is higher, at 18.1%.

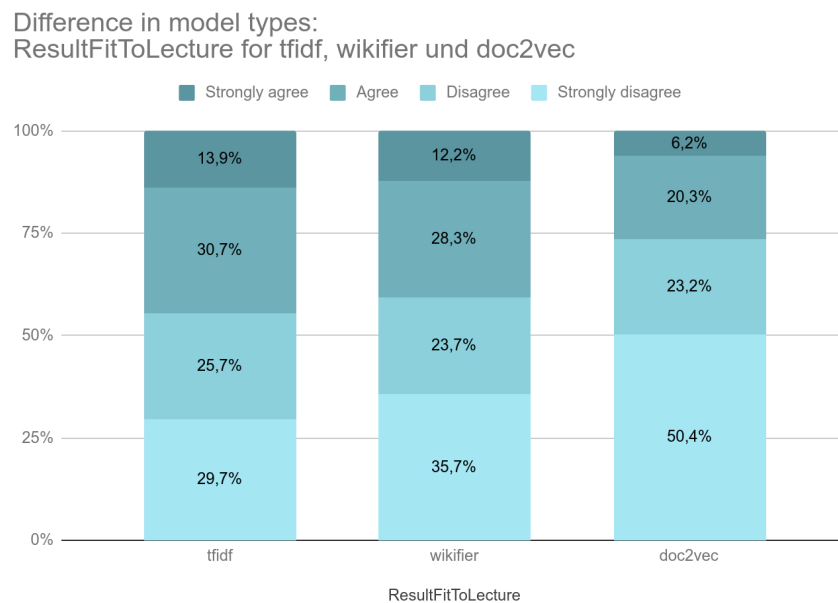


Figure 11: Bar plot of "ResultsFitToLecture" for different model types.

User comments

As described above on the web application (and technical details), users were asked to provide comments regarding the displayed OER. In total, users made $n = 62$ comments for material that caught their attention in some way. All user comments are shown in Table 12. After analysis, they were clustered into eight different categories. Some were assigned to several clusters if the user noted several issues. The chosen categories and assignment are shown on the right side of Table 12.

Summarizing, the percentage of user comments in each cluster category is:

- 52,4% “Missing course relation” ($n = 43$).
- 17,1% “No learning content” ($n = 14$).
- 8,5% “No fitting or relevant content” ($n = 7$).
- 7,3% “Document quality issues” ($n = 6$).
- 6,1% “No content at all” ($n = 5$).
- 3,7% “Only indirect relation to topic” ($n = 3$).
- 3,7% “GDPR issue” ($n = 3$).
- 1,2% “Wrong link” ($n = 1$).

4.2.4 Discussion

This is the second pilot test in which the X5gon OER search engine was tested for quality assurance purposes. Unlike the first “Discovery Pilot 1” from 2019, the search engine from Nantes and not the one from Ljubljana (JSI) was tested here. A comparison of the two pilots is possible, because both search engines are based on the same indexed data or the OER database. Also with regard to comparability, the same test set of courses was used by the University of Osnabrück from the 2019 summer semester. In addition to the overall result quality, this pilot focused on the comparison of the different data models (“tfidf”, “wikifier” and “doc2vec”) which are the basis for weighting the search results.

Since the “Discovery Pilot 1” in May 2019, the total number of indexed OER has increased by +26.82% or 23680 items and now contains 111975 items. As with pilot 1, most of ~76% of the indexed OER is in English. In addition to minor changes in the language structure of the content, more Slovenian and German OER could be indexed.

Looking at the ResultsFitToLecture variable, an improvement of the mean since Pilot 1 can be seen. For this pilot the mean was 2.06 ± 1.03 and for the first pilot 1.72 ± 0.937 on a scale of 1 to 4 according to Likert. This is an improvement of the user rating by ~20%.

Similar to the first pilot, users are more confident in evaluating whether the search results shown fit the displayed lecture in the matching segment (“Strongly agree”) and in the unfitting segment (“Strongly disagree”) if the variable HowSureAreYou is also evaluated. It is also noticeable that users are more uncertain about their evaluation if it is in the medium range.

To compare the three model types, the user ratings of the variable ResultFitToLecture were summed up with the ratings “Strongly agree” and “Agree”. In the comparison, the results weighted with a tfidf approach are in the lead with 44.6% agreement. The weighting per “wikifier” comes second with 40.5% followed by “doc2vec” with 26.5%. This shows that results that are weighted by tfidf best match the lectures according to user ratings, whereas “wikifier” results were rated only slightly lower.

This pilot had more user comments than the first pilot. The evaluation of the comments shows a wide range of previously unconsidered deficits in the quality of the indexed content. In the first place there is the wish of the user for a link to the parent course element and not only to a partial part like e.g. a document. One user discovered indexed OER that contain no educational content at all. In addition, material was discovered which, although it matched the lecture in terms of title, did

Levels / Comments	Counts	% of Total	Cumulative %	Missing course relation	Document quality issues	No fitting or relevant content	No learning content	Only indirect relation to topic	GDPR issue	No content at all	Wrong link
A complete course would be helpful	1	1.6 %	1.6 %	1	0	0	0	0	0	0	0
A strange document.	1	1.6 %	3.2 %	0	1	0	0	0	0	0	0
Again a part of a course with not so relevant content. But the course probably fits very well	1	1.6 %	4.8 %	1	0	1	0	0	0	0	0
Another resource in the same course. The course is for sure relevant. I don't want to see these parts only	1	1.6 %	6.5 %	1	0	0	0	0	0	0	0
Complete course would be more helpful	2	3.2 %	9.7 %	2	0	0	0	0	0	0	0
Course involves some basic statistics, but not much.	1	1.6 %	11.3 %	1	0	0	0	0	0	0	0
Der Titel sagt mir nichts, sehe daher keinen Zusammenhang	1	1.6 %	12.9 %	0	0	1	0	0	0	0	0
Description fits to lecture, but material has no learning content	1	1.6 %	14.5 %	0	1	0	1	0	0	0	0
Does not seem like educational content. Hi Mitja! Hi Marco!	1	1.6 %	16.1 %	0	0	1	1	0	0	0	0
Might be background material but not directly on any of the course's topics.	1	1.6 %	17.7 %	0	0	0	0	1	0	0	0
No learning content	1	1.6 %	19.4 %	0	0	0	1	0	0	0	0
No learning resource but a list of grades/points.	1	1.6 %	21.0 %	0	0	0	1	0	1	0	0
No usefull content at all	1	1.6 %	22.6 %	0	0	1	1	0	0	1	0
Not the course but the literature list is presented	1	1.6 %	24.2 %	1	0	0	1	0	0	0	0
Only a course description not a real OER resource	1	1.6 %	25.8 %	0	0	0	1	0	0	0	0
Only an uninteresting part of a course. The course may be relevant, cannot tell by this content	1	1.6 %	27.4 %	1	0	0	1	0	0	0	0
The course is about variants of the Spanish language. The variants typically also include phonetic and phonological differences so the recommended document would be relevant. But this course in particular, as indicated by the title, does not cover phonology but only style and expression.	1	1.6 %	29.0 %	0	0	1	0	1	0	0	0
The recommended material is useful as background material for the course but it does not cover topics of the course.	1	1.6 %	30.6 %	0	0	0	0	1	0	0	0
The video is not about "Human-Computer Interaction In Information Society", it's about the organizational background of a conference of that name.	1	1.6 %	32.3 %	0	1	0	1	0	0	0	0
The whole course would be better than only the quiz, as an OER entity	1	1.6 %	33.9 %	1	0	0	0	0	0	0	0
This has no context. not a valid learning resource	1	1.6 %	35.5 %	0	1	1	1	0	0	1	0
This seems to be a list of exam results. Probably even a GDPR violation. For sure not OER content.	1	1.6 %	37.1 %	0	1	1	1	0	1	0	0
complete course would be more helpful	32	51.6 %	88.7 %	32	0	0	0	0	0	0	0
leeres Ergebnis	2	3.2 %	91.9 %	0	0	0	0	0	0	2	0
link goes to a sub category of a course	1	1.6 %	93.5 %	0	0	0	0	0	0	1	1
no content, just an index	1	1.6 %	95.2 %	1	0	0	1	0	0	0	0
the better document compare to the other from the same course	1	1.6 %	96.8 %	0	1	0	0	0	0	0	0
just a linklist	1	1.6 %	98.4 %	1	0	0	1	0	0	0	0
Exam results GDPR issue	1	1.6 %	100.0 %	0	0	0	1	0	1	0	0
82				43	6	7	14	3	3	5	1
100,0%				52,4%	7,3%	8,5%	17,1%	3,7%	3,7%	6,1%	1,2%

Figure 12: Overview and clustering of user comments.

not contain any suitable or relevant content. Users commented material that contained no content, wrong or dead links to content and poor quality documents. Surprisingly, the evaluators also found documents containing matriculation numbers with corresponding grades. It can be assumed that this material was uploaded and published by mistake by a repository.

4.2.5 Conclusion

Since the first pilot, the performance of the search engine has been greatly improved in terms of result quality. This is due to a better understanding of user requirements, which were identified and implemented by the developers when the first pilot was completed. In addition, it was possible to connect new repositories to the X5gon network and to index their contents, which in turn leads to a larger selection of contents and thus to more appropriate results.

Through the user comments of this pilot we see that the data quality of the OER database still offers potential for improvement. In order to make the services developed by X5gon useful for users and other providers in the long-term, we have to work urgently on improving the data quality of the indexed OER. Based on the data from this pilot, we make the following recommendations:

- Improving the data quality of the already indexed OER with regard to the issues mentioned in this documentation.
- Developing suitable filters when harvesting new OER and repositories.
- Indexing more OER content to cover a wider range of topics and thus provide better results.
- Use of the model types “tfidf” and “wikifier”.
- Provide users with a link to the parent course (see user comments).

5 Other pilots

Although the work plan is focused on three official pilots described in previous sections, other X5gon partners with OER resources, as is the case of UCL and Nantes, have run other minor pilots to test X5gon services and tools. In addition to project partners, external organisations are welcome to pilot X5gon services and tools, not only as we do in the official pilots, but also in other innovative ways showing the value of X5gon developments. This is the case of the Kobi app in Y3, described below.

5.1 Kobi app: helping children learn to read

The creators of Kobi, an Slovenian mobile app that focuses on helping children with reading disorders, contacted X5gon for studying possible collaborations. They were interested in using some of the X5gon tools to enhance their user experience. In particular, they wanted to assess if the X5gon Transcription and Translation Platform (part of the X5oerfeed component) would be of help in detecting reading difficulties from children speech, and thus identifying the particular word(s) requiring more reading practice.

To that end, we agreed to prepare a specific web interface that, having both the predicted text and the original reference text available, shows the edit distance (insertions, substitutions and deletions) in different colors to highlight possible errors detected by the automatic speech recognition systems. Kobi creators tested the proposed solution and reported accuracy detection rates of 75% and 60% for two different test audios recorded by children with reading difficulties.

6 User studies

6.1 Introduction to X5Learn

We have developed a new platform X5Learn for accessing educational videos, and selecting additional material. The educational videos are provided in different languages including English, French and Slovak. Videos have unique ways to convey information, and engage the YouTube generation. They are at the centre and a fundamental element of the flipped classroom and online learning. However, using videos for teaching and learning presents some specific challenges. First, it can be time consuming and quite frustrating to find the required content, as video users must access the content sequentially, without being able to predict which part of the content might be relevant to their needs. Second, interesting information might be spread over several videos, or videos might be too long to fit within a classroom duration, so that teachers have to select clips. Thus, within X5Learn, we have developed a set innovative features, to facilitate interaction, information seeking and so enhance teaching and learning. We have initiated a series of pilot studies first to assess usability of the tools and user satisfaction, as well as, plan further studies to evaluate the impacts of X5Learn design.

6.2 Overview of the X5Learn User Interface

As the focus of X5Learn is on videos, the basic interface is built as a simple videoplayer model. With a search engine on the left-hand side, and resulting video thumbnails displayed in the main window. From the thumbnails, an enhanced videoplayer open for users to look at and watch videos. A number of innovative and advanced interaction features are integrated to the platform:

- a search engine: ongoing development would for example let the users search by media types and languages.
- a content flowbar from keywords extracted from the video content to facilitate browsing.
- Views: different way of looking at the videos thumbnails, including the traditional picture and different visualisations of global keywords
- A playlist: to create a playlist of videos and ways to edit them.
- Annotation and Review tool: to add video reviews and notes associated to each videos to enhance self-learning.

To make X5Learn interface intuitive to use, our user interface designs leverage familiar patterns and techniques, such as cards, popups, cascading menus, playlists, timelines and so on. To facilitate video browsing, we have developed a novel interaction technique, the Content Flow Bar (CFB) providing semantic “snippets” related to the video content that pop up on the screen as part of a video content flow visualisation (see Figure 13). This type of cueing is intended to enable the user to see at a glance what a video lecture covers and to be able to stop at particular points to discover more. As part of the CFB feature, users can look at keywords definitions, which are extracted from Wikipedia.

We have also provided additional ways for users to quickly look at the video content, instead of the thumbnail view, the user can choose between different views such as between bubble or SwimLane. Users can quickly look at the most important video keywords associated with the whole video content. Keywords are generated from the video script and represented using a timeline. On the timeline, each dot represents one of the keywords along with its associated text snippet extracted from the video content.

Video reviews or comments are a common feature of many videoplayers. They provide an additional source of information about the video, especially of its quality and interest for viewers. Thus such a feature was integrated in X5Learn. Advance features to display and aggregate reviews in different ways have been conceptualised as wordcloud and bubble, but not yet fully implemented.



Figure 13: Hovering over a video fragment within the Content FlowBar opens a popup with the semantic snippets. The Content FlowBar allows fluid preview, recap and navigation within the video content.



Figure 14: Swimlane view from video thumbnails, each node is associated with the relevant video script snippets

Some videos in the X5Learn platform can be as long as 90 minutes and as highlighted by X1 teachers rarely used such long video in their class, as it would preclude other pedagogical activated associated with the video during the class time. Thus, we added a feature to support teaching that enables users to select clips from videos.

One the main intended user group is teachers (and their students). Looking at the literature on teaching with videos [9], and complemented by feedback collected from our iterative design process, we realised another important feature in X5Learn was a playlist. Indeed, teachers who used videos as references in their classes would want prepare a global playlist for the course. The playlist feature is also useful to publish or show all video clips associated with a class. Students can also produce a playlist for their coursework.

To facilitate teaching and learning, we also introduce a note taking feature, in which notes are associated with specific segments of a video. Note taking while watching the videos can enhance students self-learning. However, notes but can also be aggregated and downloaded to be shared between students or integrated to coursework. Teachers can also share the notes with their students.

6.3 Iterative Design and User testing

Iterative design is a design methodology based on a cyclic process of prototyping, testing, analyzing, and refining a product or process (Wikipedia). The design of X5Learn follows an iterative process. New interface designs were tested by a few users, at different points in the development process. Major usability issues were fixed and the interface revisited until problems were solved.

6.4 Content Flow Bar Pilot Study

Our goal is to conduct a user study to evaluate the effectiveness of our tool and its impact. We want to understand how the CFB supports information retrieval and facilitate content navigation by examining participants video navigation and browsing patterns. The goal of the pilot study is thus twofold: to test and assess the experimental design, and insuring that major usability issues would

not introduce bias in the main study. Although an iterative design process was used while developing the tool, we also wanted more generally to highlight remaining usability issues.

6.4.1 Methodology

The pilot study consisted in comparing the performance of an enhanced videoplayer with the new content lookup tool, to a baseline video player. This kind of study is not uncommon for assessing innovative videoplayer interface designs [10, 11]. Thus, the study is based on a repeated measures design, so each participant goes through the control and treatment conditions, thus using the baseline and enhanced videoplayers. We used counterbalancing to address training effect and fatigue. The information seeking task at the heart of the study consists in finding relevant video clips that would be used in teaching. To inform the task we developed two scenarios relating to two themes: In one condition, the participants would look at videos related to Machine Learning and in the other Climate Change.

- Scenario 1 for Climate Change is as followed:

You have been asked by a sick colleague to help him prepare a lecture on climate change, and find interesting videos that illustrate how climate change can be mitigated by sustainable development, which will serve to initiate a class debate. So your are going to select 3 videos for his class. As watching the videos should not take the whole class time, you will have to select the relevant segment video segments that students will have to watch.

- Scenario 2 for Machine Learning is as followed:

Some students wanted to hear some more about the applications of machine learning in your next workshop. As it is interesting, you are going to find 3 videos that the students can watch at home, and make notes. The main topics and issues will then be review during your class. As watching the videos should not take too long, you will have to select the specific video segments that students will have to watch.

In one condition, the participants would look at videos related to Machine Learning and in the other Climate Change. We selected 18 educational videos available from the X5Learn platform for each theme. The study was implemented in a specific area of X5Learn, and participants' interactions with the platforms were captured in a log.

The System Usability Scale (SUS) was implemented more specifically in this pilot study to assess usability. SUS has become an industry standard and it is used in a variety of applications and domains. It is well documented with over 1000 publications. It consists of a 10 item rating scale with five responses ranging from strongly agree to strongly disagree (see results below).

Additionally, to gather more insights about participants experiences with the tool, we conducted short interviews and thus developed an interview guide. The main questions are as followed:

- Thinking about your experience with the tool, what was the most salient feature?
- What role did the content flowbar play in fulfilling the task? in finding information?
- Was the information provided with the videos useful? In which way?
- How easy was it to select video segments?
- In which context, would the Content FlowBar used?
- Any suggestions to improve the tool? Any additional content?



6.4.2 Procedure for the study

All the participants were sent the information sheet and consent form so it could be signed before the session, as per ethics requirement. The session took place remotely through Zoom and was recorded, the X5Learn platform was used to conduct the study.

The researcher first gave participants a brief overview of the study, then demo the X5Learn platform, highlighting first CFB and then other X5Learn features needed for the study (e.g. how to select video clip). The participants were told to review the features and practice until they felt confident that they could move to the main study. It usually took 5-10 minutes to do this, any questions about the platforms were then answered.

Then, each participant performs the information seeking task for a theme according to the relevant scenario, and for a videoplayer mode, and then for the other theme and videoplayer mode. At the end of the session, where both tasks have been performed, participants were asked to fill a usability questionnaire (SUS), which was followed by a short interview. There were 6 participants in the pilot study mostly UCLIC PhD students but also university lecturers.

6.4.3 Results

Usability (SUS)

Looking at the participants' scores on the rating scales (see Figure 15), we can observe that they are not homogeneous and vary between questions but tend toward the positive side of the scales. Most participants thought that the tool was straightforward to use and that they would like to use it. They felt that they could use the tool confidently. However, participants also found the tool was not very well integrated.

The overall usability performance in the aspects of effectiveness, efficiency, and overall ease of use are normally calculated from the ratings but the sample of participants is too low to obtain any meaningful result. The overall usability performance in the aspects of effectiveness, efficiency, and overall ease of use is normally calculated from the ratings but the sample of participants is too low to obtain any meaningful result.

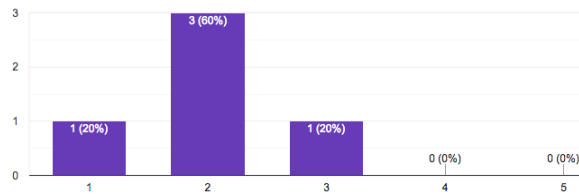
It has to be noted that one participant did not seem to like the tool and gave very negative answers. As the participant's answers were completely outside of the ratings range of other participants, thus this participant was excluded from the study, and answers removed from SUS. Results of SUS by questions, are presented in the questionnaire order, all the Likert scales range from strongly agree to strongly disagree.

Usability issues observed during the session

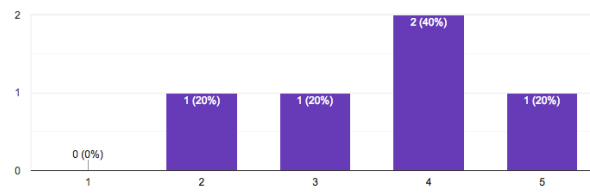
There were some usability issues occurring during the sessions, mostly with the computer itself. Some participants using a laptop had difficulties in selecting video segments with a track pad (i.e. the timeline selecting areas is quite small). A participant could not see the last keywords definitions (below the screen). Furthermore, mostly on a laptop or very old computer, some screen resolutions did not work, the videoplayer would not open. Thus, we reviewed screen resolutions and for now produce a document with best and possible screen resolution dimensions.

User Experiences

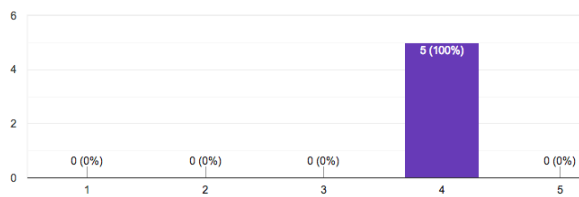
When analysing very broadly participants' interviews, we can see a number of broad themes emerging. Participants liked the Content Flowbar, they find the function appealing and could see its value for information seeking: "I quite like the idea of having an easier way of kind of scanning through



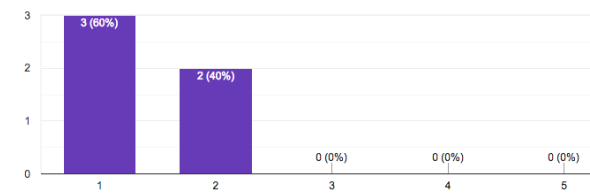
(a) I think that I would like to use this system frequently.



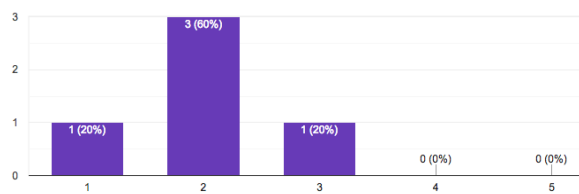
(b) I thought there was too much inconsistency in this system.



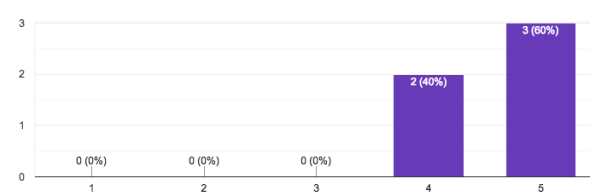
(c) I found the system unnecessarily complex.



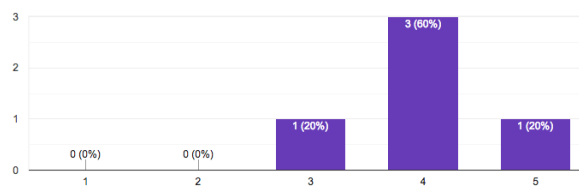
(d) I would imagine that most people would learn to use this system very quickly.



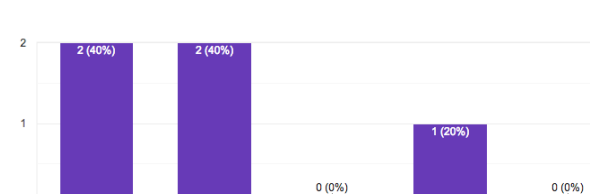
(e) I thought the system was easy to use.



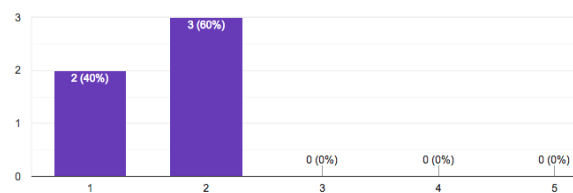
(f) I found the system very cumbersome to use.



(g) I think that I would need the support of a technical person to be able to use this system.



(h) I felt very confident using the system.



(i) I found the various functions in this system were well integrated.

Figure 15: Usability (SUS) results

video and getting an overview of the video.” and “Seeing how the videos were segmented, I think from experiencing both, it was evident that it was a lot easier to navigate the long videos using the keywords.”

Despite being very positive about the CFB, participants were wondering about the relevance and accuracy of the keywords, with outliers such as “Bolognese sausage” and “Donkey”. Moreover, they also question the pertinence of some keywords such as Machine Learning, as this study all videos pertain to Machine Learning, or having the same keyword like Global Warning in every segment of the video. An interesting observation was made by a student, she said she was afraid while using the CFB to become too reliant on the keywords, and missing engaging part of the video. Participants in this sample did not need the keyword definitions, but thought it could be valuable to undergraduates, as videos can contain very technical terms. One participant noted that it would be more interesting if keywords could be more context sensitive.

Participants thought that the tool was quite intuitive to use, but some participants stated that would need to familiarise themselves somewhat with the tool. As a participant commented, “I think it works, it needs to get a bit of time to get used to but it’s pretty easy to use it”. Several participants mentioned that some videos used in the study were quite old, so they would not use them in teaching as their relevance was quite questionable.

Last there were quite a few suggestions on how CFB could be used in different teaching contexts from lectures to workshops, and art projects. Besides teaching, two interesting propositions related to conferences and hobbies. Indeed in the present context, most conferences take place online and often recorded. Thus pointers and clues for what to watch could be very useful, as users have limited time-frame. As suggested, the tool could also be very helpful for personal video collections related to hobbies, CFB could thus assist users with large and not well-maintained collections. Further studies should thus be conducted to disseminate CFB in other domains.

Revisions to study

With some cautions with the equipment that participants will be using, there seems that no major issues would preclude conducting the main study. From the participants interviews, it was decided in the main study to look in more depth at some attributes and effects of the CFB. Thus the research instrument for the study was adjusted using new rating scales pertaining to CFB. We also made every effort to remove and change the oldest videos.

6.5 Playlist Pilot Study

6.5.1 Introduction to Playlist

Playlists have become an important feature of video-based teaching and learning. They provide an external way of showing and ordering media to be played (e.g. songs, videos). Research in this context has focused largely on the commercial YouTube ‘Playlist You’. Such studies have been reported for a variety of domains including health, the Sciences, and Human-computer Interaction. One of the main topics to be covered include how best to design and present video playlists in an educational context – for both student and teacher. Very short video previews have also been created for conference attendees to give them an overview of upcoming talks [12].

A video playlist for a course can be composed of a lesson plan with specific objectives related to the playlist and several educational video clips. This can be particularly helpful for students moving into a new field of study. A further reason for designing playlists of videos with specific topics is that it can provide students with easy access to material and help them and their teachers when referencing the content conveyed in the videos. Video playlists can also assist in students or teachers structuring

background knowledge for a topic that is being covered in a course, e.g. climate change. Playlists can be annotated to enhance students' learning, which is not an easy task as teachers have to decide how best to annotate them in a particular context. Playlists can also be seen as modern version of a reading list [9]. They can motivate students to look further afield by selecting other material on the list. Playlists can also provide students with a degree of quality insurance that the videos will have relevant and well-presented material, especially when provided by teachers.

But how best to construct a playlist? Their design has been discussed within specific pedagogical practices, for example Green et al. [13], who created video playlists in the context of case-based teaching, to illustrate the case material and provide additional viewpoints. Snelson [14] has also argued that playlists can be mapped to different learning modalities, such as affective or cognitive learning to accommodate different learner groups. Playlists can also be shared, exchanged or co-created by different teachers. Playlists have also been used in the context of blended learning and the flipped classroom where students need to take more initiative in choosing the video material they watch outside of the classroom.

Another approach is to ask students themselves to make and annotate playlists. This active form of learning may help them obtain an in-depth insight into a course subject. They can remix playlists, adding new videos or clips, and produced reference playlist for essays. Some courses may contain hundreds of hours of video lectures which can be daunting when students are embarking on revision for a course. For this, teachers can create specific playlists intended to help facilitate revision by giving guidance on what to view for given topics.

An overarching question this raises is: how best to design the interface of an interactive digital playlist that can support the various learning activities outlined? Our research thus aims to answer this question providing playlist that teachers can create and use in these different contexts from our open education resource (OER) portal X5Learn. Next, we describe the functionality that was built for this purpose.

6.5.2 Playlist functionality

The playlist that was implemented in X5Learn has the following features and task steps that are also illustrated in Figures 16, 17 and 18:

- A teacher can initiate a new playlist by selecting “create a new playlist” function.
- They can then use the search engine to find appropriate teaching material.
- From the results list, a teacher can then use the interface functions called the “Content FlowBar” and “Views” to locate items of specific interest.
- Teachers can add the videos to their playlists using the video player, open videoplayer and click on add to playlist, then select for which playlist (see Figure 17).
- Once all the required videos have been added to the playlist, the playlist creator can then order the list of selected OER materials in the sequence of their choice. Or they can use a specific AI function, called “optimise learning” that is based on a Machine Learning model, which will automatically reorder the video sequences based on its assumed best learning sequence.
- Teachers can also annotate videos, for example add specific comments and questions to selected video fragment. They can dictate or type the notes, and aggregate them for example to produce a coursework sheet (see Figure 17). They can also modify the title and the description of a video (see below in the next paragraph for further description).

- When the collection and the order are finalised, a teacher can then “publish” the playlist as a new resource in the X5Learn OER repository.
- The teacher assigns the final “title”, “description”, “author/s”, and a “license” when publishing.
 - The content creator gets a confirmation email in the email registered with X5 Learn accompanied by a URL that will let any learner access the playlist.
 - The teacher / user can also download the playlist and share the resource via a link with students (see Figure 16)
- Playlist creators can also clone playlists or duplicate them, for edits and modifications, to make different playlist versions that can be published again, for example for different groups.

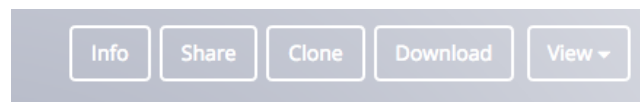


Figure 16: Playlist options after publishing

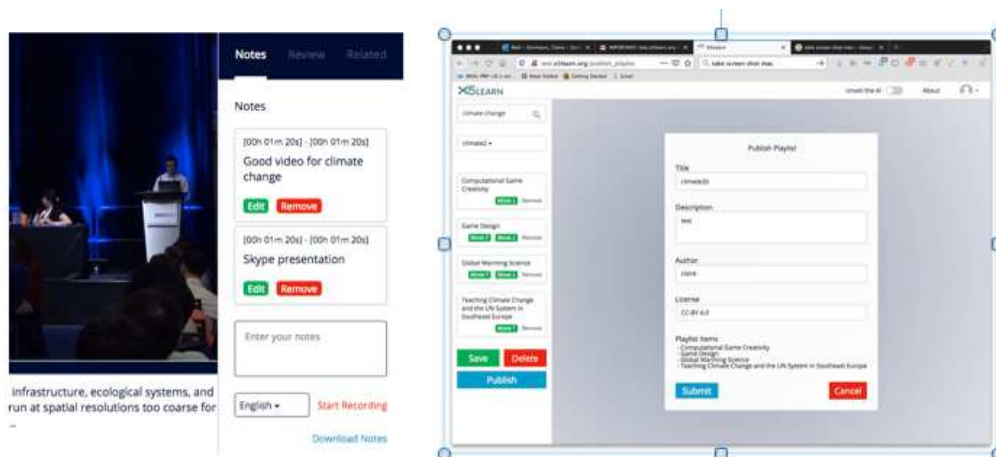


Figure 17: Playlist Creation, Annotating videos & adding video playlist

6.5.3 Designing X5Learn Playlists Tool: Initial user study

During the development of X5Learn Playlist, an iterative design process was used, and we gathered feedback from users, formally and informally, at different stages of the design process. A large number of the Open Educational Resources (OER) indexed in X5GON are university level materials. The current remote teaching setting in universities gave us the opportunity to test the playlist creation tool with a few university lecturers. In particular, two university lecturers affiliated to University of Peradeniya, Sri Lanka, teaching “Computer Science” and “Agricultural Biology” run initial assessments of the playlist feature.

They were asked to create a playlist that they could use in their teaching, and then short interviews were conducted to understand their experience. The interviews focused on:

1. What they liked.
2. The challenges they faced.
3. How the challenges could have been overcome.

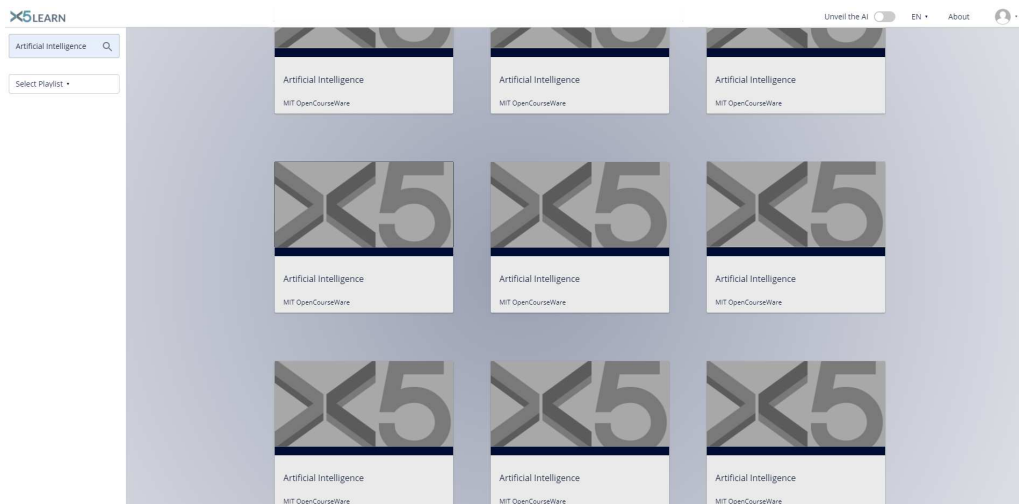


Figure 18: Search result for “Artificial Intelligence” in X5Learn initial interface

Interviews with University Lecturers

At this point, the X5Learn design was in its early stages. The two participants were briefed on how the system works and provided with a User Manual. Then they were instructed to try out the X5Learn system while trying to create an OER playlist that they can incorporate in their teaching.

The OER resources that are currently available via X5Learn on Agricultural Biology were found to be limited making it difficult for the lecturer in Agricultural Biology. She had a real struggle to find relevant resources in that could be incorporated in the study programme: “I can see myself using many features that are available in the playlist tool and saving a lot of time. But, at this point, I can hardly find any materials that I can use in my courses. Maybe, at a later stage when the materials are accessible, I am happy to give it a go again.”

However, she appreciated the overall design and the features available in X5Learn. She said that the ability to share a web link with her class, during the COVID-19 pandemic was essential. She saw how finding diverse resources from world-renowned institutions could make her resource creation process more efficient and easier.

By contrast, the lecturer who taught computer science had plenty of resources in X5 Learn that could be used, especially for the topic of Artificial Intelligence: “I’m searching for ”Artificial Intelligence” ... since I’m teaching the course these days...”

Hence, their experience of creating a playlist about “Artificial Intelligence” provided more insights about the X5Learn playlist tool. For this lecturer, videos were not enough to make a good playlist. They also suggested including PDFs to complement videos. He also pointed out the usefulness of the ordering feature in the playlist tool: “... The ability to reorder content once the materials are selected is very useful. Usually, I don’t find the materials I want in the order I want my students to look at them ...”

However according to this lecturer, there were several issues with the tool. The main challenge was the difficulty in find the right material quickly. At this stage, the title and the source were the only information provided with the material (as seen in Figure 18). Many videos that were extracted from the same source tend to have the same title in the meta-information. This phenomenon is very common in MIT Open Courseware (<https://ocw.mit.edu>) where all individual resources (PDF, videos, etc.) associated with an open course tend to have the same meta-information. Likewise, X5Learn (X5GON) assigns the same title to all the resources from that course. This means that a user has to open individual items in the video player to “preview” what the material is, which is quite

cumbersome.

Another minor challenge was the resolution of the video player was not well adapted for the PDF, as too small (see Figure 21) which makes it hard to read the contents of the material. This negatively impacts the user experience as the user has to put extra effort to read the small letters or work with a zoomed version of the content that makes navigating through the PDF files cumbersome.

Recommendations for Improvement

The two participants also provided feedback on how the system can be improved. The main recommendation was to expand the resources available via X5Learn (X5gon), for example, by identifying OER in the YouTube video repository. Another suggestion was to include a button in the video card to add materials to the playlist to reduce the number of clicks on the user's behalf. Reducing the number of actions needed to fulfill a task is a common approach to improve usability and effectiveness of computer systems: "... Ability to add to playlist from the search results window (rather than clicking and moving into the page) ... lesser number of clicks ... I might go through the items and then come back to create my playlist. Don't want to click twice to add..."

Another recommendation was for the playlist author to be able to change the title of the material: "I should have a way to rename the items added to the playlist...". This recommendation is motivated by the problem described above (in Figure 18) where many items have the same title making it hard for the user to uniquely identify learning resources. It was also identified that allowing the content description for each learning item to be modified (i.e. in video player) would allow the author to include instructions or relevant information with each item in a playlist.

Modifications of X5Learn Interface

Based on the issues and the recommendations proposed by the lecturers, several new features were added to the X5Learn interface. An additional feature was deployed to automatically generate thumbnails to the learning resources that are presented to the users as shown in Figure 19. This allows the user to have an instant preview of the learning resource before committing to investigate it thoroughly.



Figure 19: (Left) The resource cards with the X5 logo that gives no insight into the contents in contrast to (Right) the resource cards with thumbnails allowing the user to get an instant preview into what they are likely to find inside the resource.

The new version of X5Learn also allows the title and the description of an OER to be modified once it is added to a playlist as portrayed in Figure 20. This feature allows the author to tailor the title and descriptions of the items to align better with the intended learning outcomes of the playlist.

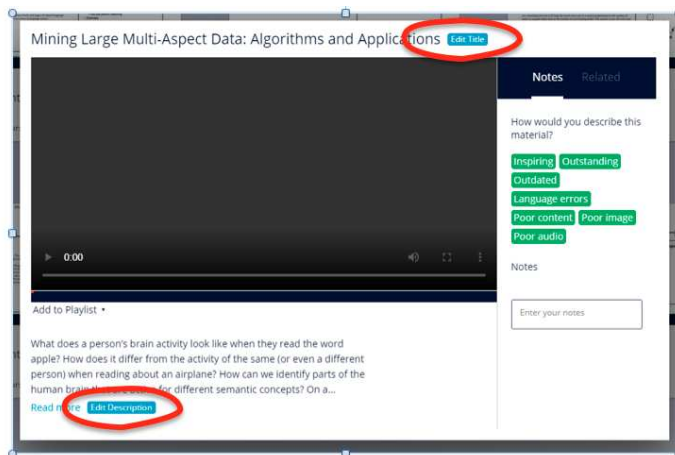


Figure 20: The two buttons circled in red, “Edit Title” and “Edit Description” that allows the title and the description of a playlist item to be modified once added to a playlist.

The video player has an inbuilt option to view the video in full screen in case the video is too small. However, having the same feature for PDF was rather challenging as the PDF reader is constrained by the pixel area that is allocated to it by the application. To overcome this challenge, a new button was added to the document viewing screen titled “Expand Document” (circled yellow in Figure 21). When clicked, this button will open the PDF document in a new web browser tab allowing the full browser space to be used for viewing the PDF document.

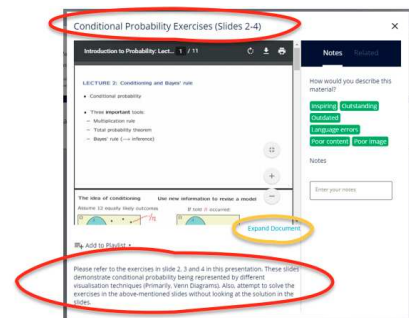


Figure 21: Edit title and description view

6.5.4 Pilot study

Goals of the Pilot Study

The main goal of the pilot study was to assess the usability of some of X5Learn platform features, and more specifically for the playlist. We wanted to examine in more depth the type and severity of usability issues, and gain feedback to refine the feature design. According to Hagen [15]: “the user asks the same sorts of things from the streaming service: ease of use, accessibility of content, and an overview functionality that is effective and comfortable. Playlists become fixed entities in a technology defined, ultimately, by its fluidity”

The second aim was to gather some insights on the behaviour of high-school teachers regarding video information retrieval, and their experience with using the educational platform X5Learn. High-school teachers and their students are quite a different target group than university teachers, and X5Learn has potential to support this user group.

Participants

We decided to focus on maths instructors to evaluate the tool for math teachers’ training. We recruited 5 teacher trainees for our pilot study. All of the participants had some exposure to teaching and had taught at GCSE level in the classroom.

Selecting Videos for the Study

At the time of the study, X5gon did not contain any math videos for high-school students. We thus selected videos outside of X5gon, which were ported to X5Learn. We selected videos from 3 sources: from a MIT series dedicated to high school students, and two from YouTube video series dedicated to GCSE math exams. The MIT series included the most advanced material ¹.

One of the YouTube followed a common design strategy in this domain, using white boards for writing down formula and examples combined with a running commentary ².

The other video series was chosen as it was designed and includes features to engage students ³. At the time of the study, the interface was only partially functional; “optimise learning” and “download playlist” were not fully functional.

Scenario and Study Tasks

To assess the playlist, we created tasks that had to be completed by the participant teacher trainees based on the playlist functionalities. Thus, we developed this scenario to guide participants in the study: “You have a class with students of mix abilities. Some of your students are still struggling with their GCSE exam preparation, while others will take an A level in math. To help weaker students with their exams, you are creating a playlist list with two videos that they should watch at home, before doing exams practices. One video should cover negative numbers, and the other a topic that your students struggle with.”

For the most advanced students, the participants were then asked to add two videos that demonstrated how to bridge GCSE to A level. Then they will get a different set of exercises to do. During the class, you will show some highlights of the videos and review them with students, before doing exam questions or relevant exercises.

Think-Aloud Method

An adapted version of the think-aloud method was used. This consisted of observing a user working with an interface while encouraging them to “think-aloud”. It is useful to capture users experience while they are interacting with a design [16] and for assessing usability aspects of a new interface. The pilot study was thus centred on using a think-aloud protocol. Following this method, participants were given the standard instructions ⁴: “During this study, I will ask you to verbalise your thoughts as you are using X5Learn to make a playlist, while you are performing tasks such as selecting videos, adding them to playlist and publishing it. Tell me, please, what you are looking at, thinking, doing, and feeling as you go about your task. Some examples of think-aloud statements are as follows:”

- “I want to do...”
- “I’m looking at the navigator screen, and I think it does...”
- “Hmm, that’s not what I expected; I thought it was going to...”
- “That work well”

To complement and gain further insights in the participants’ experiences, short interviews after the study were conducted. First to gain some clarifications, participants were probed around 2–3 main issues that were observed during the think-aloud sessions. Then, we asked the following questions:

¹https://ocw.mit.edu/resources/res-18-005-highlights-of-calculus-spring-2010/highlights_of_calculus/big-picture

²<https://www.youtube.com/watch?v=tlKN8NNNxdI>

³https://www.youtube.com/watch?v=izQGGG_5rAE

⁴<http://predictivemedicine.northwell.edu/usability-lab/think-aloud/>

- When you look back to your experience what was the most salient feature of the X5Learn platform?
- Was the playlist feature helpful? (why / why not)
- Was the information provided with the videos useful? In which way?
- What role did the keywords play?
- Do you have any suggestions for improving X5Learn?

Protocol for the study

We sent the study's information sheet and consent form to participants prior to their sessions, and we recorded the Zoom's sessions. We gave participants the scenario and instructions for the think-aloud. We then demoed X5Learn main features, and let participants practice for a short while before the main study. At the end of the study, participants were probed around 2–3 issues before moving to the interviews. It has to be noted that to simplify the pilot study and make it shorter, we did not include the search engine in it.

6.5.5 Results

Overall, participants generally found the interface easy to use. A number of usability issues were uncovered during the pilot study. These are described broadly by features.

Keywords and Associated

Some participants used or looked at the view feature during the study. One selected “bubble” instead of thumbnails, and the other as a complementary means to visualise keywords. Although the participant who tried them all did not understand the differences between the views and what the different visualization representations meant. He also did not realise that dots in the swimlanes visualisation were associated with snippets from the video content. This suggests that a novel interface like this one requires a way of helping the users learn quickly what the different features mean.

All participants mentioned that numerous keywords were both useful but sometimes difficult to use in this context. As one participant stated, keywords were a double edge sword: when it worked it worked well, but when it did not, you just end-up with random keywords. Another participant mentioned that keyword definitions were quite interesting and would be useful for students, as they could read definitions while watching a video. They suggested it could help students to understand the videos better and thus support learning.

One participant initially thought that the definitions were part of the video, and thus was trying to select a video segment accordingly, but then realised that this was not the case. Two of participants mentioned that they would have like the retrieved videos to be better organised, by topics or categories.

The Playlist

Creating a playlist and adding videos to it was found to be relatively straight forward. However, some participants, at first, did not seem to realise that they had to go back to the “create playlist” and select their list to be able to visualise it.

One participant did not initially grasp the concept of playlist. She started to create a list for each of the selected videos, and upon realising how it worked then tried to publish the first one. The message “need more than one video appeared”. Her mental model was based on Google Drive where you uploaded one video at a time. Thereafter, she mentioned she liked the idea of having this different set-up.



We did not tell the participants that “optimise learning” was not working at the time of the study, but did so if they tried to use it, explaining the problems. A couple of participants tried to use the feature and then asked what it was doing. Three participants used the description in the publish form to put some information for their students, although one asked if that was the intended used.

After publishing their playlist, one participant discussed at more length the playlist features. He did not seem to understand the logic behind “publish”, you cannot directly edit the playlist but have to clone it first. It seemed over complicated.

User Experience

During the study, it became apparent that participants relied heavily on titles and thumbnails to select videos. One participant mentioned using video-style, for example, to assess/guess if a video would be suitable for which math level. Participants seemed quite reluctant to watch the videos in full, instead flicking through them, looking for specific visual cues to assess the suitability of videos, such as the form and parameters of equations. Two participants indicated that they would not have time to watch whole video chunks.

The participants also commented on how they hardly ever use videos in their classrooms but might use videos for GCSE exam preparations. Some participants specified that they occasionally looked for pedagogical material in videos for preparing their class but would not use them in it. They tend to use material posted on the school internal network or turn to information suppliers provided by their schools. Nevertheless, they all agreed that free access to material would be very welcome as schools do not have big budgets, but it would need to follow school standards.

Design recommendations

Rather few usability issues were identified. Table 9 presents these and suggested improvements.

Usability issues	Solutions & Improvements
Optimising Learning	<ul style="list-style-type: none"> • Additional help messages to make function more visible. • Undo feature could be considered, if users did not like the optimisation, they should be able to return to the previous state.
View	<ul style="list-style-type: none"> • Add information (help) about different types of view. • Show an example of dot messages by default.
Publish Form -Description	Short message to state what and for who description is intended.
Help messages	More generally, small information messages describing some functions would assist users.
New User	Short video: demo how to get started.
Search results	<ul style="list-style-type: none"> • Make users aware of how results have been displayed / categorized. • Let users refine according to some criteria (e.g. year).
Edit Publishing	More user testing needed

Table 9: Usability issues and solutions.

6.5.6 Discussion

Overall, there were several positive comments about the potential of X5Learn and the playlist functionality. The participants reacted favourably to the tool, thinking about how they could use it for making playlists for home use. Some of these could be solved simply by providing short help/information messages. For example, “optimising learning” could include a pop-up window saying, “ordering the playlist video sequence to support learning”. To help new users get started, providing a short video would be helpful. Further attention should be paid to editing playlist, as it could be quite confusing.

At first, it might seem that maths trainee teachers in traditional high schools may not be the obvious target group as currently they don’t use videos, due to timetable and time pressure, or lack of suitable equipment. However, this could be overcome with providing additional relevant content in X5Learn so they are given a new resource for supporting their teaching. This could be especially important when switching to online learning where such resources could be blended with their traditional methods. Recently, Kalinec-Craig [17] showed the usefulness of using video playlists with math trainee teachers during their studies seeing teaching strategies in action. Doing so, could also stimulate the use of playlists as a math teaching tool. Furthermore, the situation could be quite different for A level, of math as extra-curriculum activities. Some of the A level math teachers mentioned they were always looking for examples to demonstrate the use of math in the real-world, and which could link A level maths to professions and university studies. They could make playlists of such videos to be used in classes or to be viewed at home.

Parents and students could also be another target group for using our tool to create playlists of online videos, especially at exam times (such as GCSE or A level in the UK). In that case, with relevant content, providers, teachers and students could make good use of the repository such as X5Learn and the set of tools we provide.

6.6 Conclusion: X5Learn

We have advanced a new platform X5Learn for accessing educational videos, as well as, complementary pedagogical material in text format. To enhance teaching and learning and support users, we have developed a set of innovative features. One of these features, the Content Flow Bar, facilitates information seeking by letting users browse through video content. Related mechanisms have been provided so users can get different “views” of the video content. Another major feature of X5Learn is the playlist. User testing of the playlist was found to enhance learning by enabling users to tailor and add specific information to each video. The playlist was designed to be versatile; it can also be integrated to other educational platforms such as Moodle (as well as other X5Learn/X5gon attributes).

Pilot studies related to the Content Flow Bar and Playlist have shown that the X5Learn interface was intuitive and easy to learn. Participants liked the features and found them valuable. At the time of the studies, there were still a number of minor usability issues to resolve. The Content Flow Bar was still being improved. A full study of the Content Flow Bar has subsequently been conducted to look more closely at users interaction and navigation patterns.

A participant in one of our pilot studies suggested an innovative utilisation of playlists, for students (and teachers) might also be in the Arts and Media to showcase their work. For example, students could make playlists highlighting specific features of their work that examiners should look at more specifically. Applications of the Content Flow Bar was also mentioned for personal collections and hobbies.

To the extent that playlist provide an opportunity for users actually to implement the state of their thinking at a particular time, it is not impossible to believe that flowbar-like tools might provide more general support in educational contexts. Of course this would involve a much more focused investigation into the iterative design of the tools, and corresponding pedagogic strategies.



Besides use cases, further development of X5Learn and its tools can be envisaged such as the integration and visualisation of information (e.g. notes and reviews). Moreover, the implementation of translation features were discussed but only partially implemented. Such tools could be interesting in bilingual courses, or when students of different languages collaborate in group works, etc.

7 Advanced cross-lingual and cross-modal features

As discussed in the introduction, the second main subtask of Task 5.2 is to pilot advanced cross-lingual and cross-modal features. This subtask, led by UPV, has been divided into four research lines: streaming automatic speech recognition (ASR), simultaneous machine translation (MT), multilingual MT, and cross-lingual text-to-speech (TTS) dubbing. Although each of these lines is of great interest on its own, it is clear that an accurate pipeline “streaming ASR → (multilingual) simultaneous MT → cross-lingual TTS dubbing” for (offline and) *live speech-to-speech* translation of audio streams would have an immense applicability in our everyday life and, particularly in Education. In the X5gon case, the kind of relevant applications we have in mind ranges from *fast* speech-to-speech translation of prerecorded (educational) videos (seen as a simple form of live audio sources) to multimodal annotation of OER (e.g. with learners’ comments), and live multilingual speech-to-speech translation on an educational platform (e.g. X5Learn), either for live lecturing or (live) user dialoguing (e.g. on particular OER). With this in mind, in Y3 we have pushed a number of activities along these research lines which are described below for each line separately.

7.1 Streaming ASR

From the very beginning of X5gon, it was realized that we are reaching the point at which (raw) automatic transcriptions are often good enough for direct publication in many cases. In particular, with a WER of 11.7% in M0, this was clearly the case of Spanish ASR for poliMedia (Section 3). Convinced that, over time, ASR progress would only make this clearer, we began to study how best ASR systems can be adapted to the so-called *streaming setup*; that is, subject to the constraint that output must be delivered in nearly real time, only within a short delay (latency) after the incoming audio stream. Needless to say, accurate streaming ASR not only would allow fast transcription of educational videos but, as discussed above, it would certainly open the door to many other AI-based advances of high value for Education.

At the beginning of Y3 (M25) or, even better, soon after (M30), we had already confirmed that Spanish ASR for poliMedia was not the only case in a good position for ASR adaptation to the streaming setup. More precisely, in M30 we had already achieved transcription errors of only 9.1% for Spanish (poliMedia), 18.8% and 15.8% for English (in VideoLectures.Net and poliMedia, respectively) and 22% for Slovene (VideoLectures.Net) [18, 5]. Therefore, in Y3 we were in a good position to adapt our offline ASR systems for these languages to the streaming setup.

As discussed in [18, 5], our offline ASR systems follow the (conventional) hybrid approach in which two separate models, the acoustic and language models, are trained separately and then combined during search (inference). These systems (and the models they use) cannot be directly applied in the streaming setup since they require the full audio signal being available and no strict constraints on the response time. Therefore, to adapt them to the stricter streaming conditions with minimal degradation, a number of novel techniques were devised and tested. Generally speaking, we re-engineered our offline (deep neural network-based) acoustic and language models and tools for the systems to work with a window of limited duration (sliding over time) and to respond quickly, with minimal latency and a stable regime. In this context, however, we prefer not to enter into much detail on the complex models involved, which has indeed been reported in [18, 5], and the way they were adapted

and empirically assessed. The reader is referred to [19], where this work has been recently published in part, and [20], where, hopefully, all details and more comprehensive empirical results will be soon provided.

As far as X5gon is concerned, we tried our streaming-adapted M40 ASR systems for Spanish on poliMedia, English on VideoLectures.Net and poliMedia, and Slovene on VideoLectures.Net. In all cases, they were adjusted to operate smoothly with a latency of just 0.8 seconds. Table 10 shows the WER scores provided by offline and streaming-adapted M40 ASR systems on VideoLectures.Net (in English and Slovene) and poliMedia (in Spanish and English), as well as the relative WER increase of the streaming systems with respect to their baseline, offline counterparts.

Table 10: WER scores provided by offline and streaming-adapted M40 ASR systems on VideoLectures.NET (in English and Slovene) and poliMedia (in Spanish and English).

ASR system	VideoLectures.Net		poliMedia	
	En	Sl	En	Es
M40 Offline	14.8	15.3	12.0	8.3
M40 Streaming	15.4	15.8	13.4	8.7
$\Delta\%$	4.1	3.3	11.7	4.8

From the results in Table 10, it becomes clear that adapting offline systems to the streaming setup is perfectly feasible at the expense of a minor relative error increase. Indeed, with the exception of an 11.7% relative WER increase on poliMedia in English, we see that the average WER $\Delta\%$ is only around 4%.

7.2 Simultaneous MT

In deliverables D3.4 [18] and D3.5 [5], all MT systems developed and reported for X5gon worked at sentence level. They must first receive a whole sentence, and only then will the translation process start. In a real-time face-to-face communication setup, this behaviour has the obvious disadvantage of participants having to wait for some time until a complete sentence is available to the system before the translation is generated. We believe that certain learning activities, such as course co-creation, whereas two teachers speaking different language work on a collection of OER, or paired-learning between multiple students each of them communicating in their own mother tongue, would benefit from a simultaneous MT tool that would allow for almost real-time cross-lingual communication.

During X5gon, excellent results have been obtained in off-line MT tasks. The results reported in [5] show how the X5gon system obtains very competitive results compared with Google Translate. Moreover, the systems developed for X5gon significantly outperform Google Translate in language pairs that are often overlooked, such as translation between English and Slovene, as well as between Spanish and Portuguese. Thus, off-line MT system can be used as a starting point for developing simultaneous MT systems.

Recently, some variants of attention-based architectures that are able to carry out simultaneous translation have been presented [21, 22, 23, 24, 25]. Basically, these variants limit the attention mechanism to those input words available in the stream, since the complete sentence cannot be observed. In this way, partial translations can be produced without having to wait for the whole input sentence. The simultaneous MT task adds the challenge of having to balance a quality-latency trade-off. If a model is forced to have low latency, translation quality can drop so much that it becomes unfeasible to use. Conversely, high latencies can ensure almost no quality drop, but the system will not output translations in a timely fashion.

As a starting point, we trained Spanish into English (Es \rightarrow En) models following the Hard Mono-

tonic Multi-Head Attention architecture [24] (MMAH), a modification of the standard Transformer [26] architecture, wherein each encoder-decoder attention head behaves as an independent monotonic attention mechanism [24]. This architecture is trained using the standard cross-entropy loss, and an added latency term that tries to minimize the delay between attention heads. The latency is scaled by introducing a scaling term λ_{var} . At inference time, the model has a dynamic policy to decide when to write words or wait for more context.

We have also trained models following the recently proposed Multi- k framework [25], which is an evolution of the wait- k policy [22]. Under the wait- k policy, a model will first wait until the first k words have arrived, and will then alternate between reading one word and writing one word until the sentence finishes. Following the Multi- k framework, models are trained with different k values, and then, at inference time, k can be fixed for the specific needs at that time.

An evaluation of the performance of the simultaneous systems is shown in Table 11 for the Es→En models, and in Table 12 for the En→Es models. The models were evaluated using the official dev and test sets of the Workshop in Machine Translation (WMT) in 2012 and 2013, respectively, under the same setting as the off-line systems, whose results are shown as a baseline. The details of the system developed up to month 24 are reported in D3.4 [18, Section 3], and the details of the month 30 system are reported in D3.5 [5, Section 3].

Table 11: Comparison in terms of BLEU score between off-line baseline and simultaneous MT systems for Es→En translation

Model	λ_{var}/k	BLEU	
		Dev	Test
Offline M24(BASE)		34.7	32.2
Offline M30(BIG)		39.2	35.9
MMAH(BASE)	0.1	22.0	20.5
	0.2	21.0	19.9
	0.4	22.1	21.4
Multi- k (BIG)	1	24.9	22.7
	2	27.5	25.4
	4	30.4	28.5
	8	33.2	31.1
	16	33.7	32.1

Table 12: Comparison in terms of BLEU score between off-line baseline and simultaneous MT systems for En→Es translation

Model	λ_{var}/k	BLEU	
		Dev	Test
Offline M24(BASE)		35.0	32.2
Offline M30(BIG)		38.0	34.6
Multi- k (BIG)	1	28.5	25.1
	2	30.3	27.1
	4	33.7	30.3
	8	34.9	31.5
	16	35.1	31.7

As seen, the performance of the MMAH models is much lower than that of off-line models with a significant drop of more than 10 BLEU points. We also tried training MMAH BIG models, but we

discarded this idea as the performance was almost the same. However, the Multi- k models do show interesting results, as they are able to match the performance of the off-line M24 model with the most favorable latency condition, with a gap of around 3 BLEU points with respect to the improved off-line M30 model. We have selected a more realistic $k = 8$ for Es \rightarrow En, and $k = 4$ for En \rightarrow Es for inference, leaving us with a gap of 4.8 and 4.3 BLEU points, respectively.

7.3 Multilingual MT

Multilingual MT allows the deployment of systems that can translate from many-to-many languages [27, 28, 29]. There are two significant advantages behind these multilingual systems that are usually stated. The cost of maintenance and deployment is significantly reduced since a single system that translates from N into M languages can potentially replace N -by- M MT systems. In addition, low-resource languages, such as Slovenian, can benefit from being trained together with rich-resource languages as a consequence of the so-called transfer learning.

In X5gon, both advantages mentioned above were appealing and were worth testing to evaluate the feasibility of multilingual MT in the OER context. Two multilingual systems were developed and evaluated with the main purpose of improving the translation quality of language pairs involving Slovenian as a first objective, but also to compare the performance of multilingual systems with bilingual systems in M30 [5, Section 3.3]. The first multilingual system considers Slovenian as a pivot language from and into English, German, French and Spanish. The second system takes English as a source language and translates into German, French, Spanish, Italian and Slovenian. More details on system training, data resources employed, and evaluation are provided below.

To train a multilingual MT system, parallel corpora with different source and target languages can be put together, but source sentences must be prefixed with an artificial token indicating the corresponding target language for each source sentence [27].

The model architecture behind a multilingual MT system can be basically the same as that of conventional bilingual MT systems, that is, the well-known Transformer architecture [26]. However, according to [29], the accuracy of multilingual systems can be improved by training deeper, instead of wider models. For this reason we opt for a variant of the Transformer architecture based on a dynamic linear combination of layers (DLCL) that allows for deeper models, that are smaller, faster and more accurate than the well-known BIG variant of Transformer [30]. More precisely, our multilingual systems followed the DLCL (BASE) configuration with 12 layers, 512K words per batch and 16-bit floating point implemented in the Fairseq toolkit [31].

Table 13 states basic statistics of the parallel corpora devoted to training, development and test involving the first multilingual MT system that considers Slovenian as a pivot language. The training sets are a selection of OER-related parallel corpora available at the OPUS website⁵ for each language pair [32]. As observed, the total amount of sentence pairs in the multilingual training sum up to 98 millions. As defined in deliverable D3.4 [18], the in-domain development and test sets are those of the in-domain VideoLectures.NET (VL) task, while additional out-domain test sets are those publicly available at the IWSLT evaluation campaign [33].

Table 14 shows basic statistics of the English-source multilingual training corpora with over 300 millions sentence pairs devoted to train a multilingual MT system to translate from English into German, Spanish, French, Italian and Slovene. Those corpora are a super-set of those employed to train the bilingual systems reported in deliverable D3.5 [5].

Table 15 shows BLEU scores of the bilingual En \leftrightarrow Sl systems reported in deliverable D3.5 with minor updates and those achieved by the multilingual Sl \leftrightarrow {En,De,Es,Fr} systems. As seen, bilingual

⁵<http://opus.nlpl.eu>

²Backtranslations obtained using a De \rightarrow En system.



Table 13: Basic statistics (in millions) of the Slovenian-pivot multilingual training corpora

Pair	Sentences	Words		Vocabulary	
		Source	Target	Source	Target
De \leftrightarrow Sl	19	256	236	5.0	4.4
En \leftrightarrow Sl	30	390	331	6.5	6.4
Es \leftrightarrow Sl	25	296	248	5.1	4.6
Fr \leftrightarrow Sl	24	336	268	5.7	5.0

Table 14: Basic statistics (in millions) of the English-source multilingual training corpora.

Pair	Sentences	Words		Vocabulary	
		Source	Target	Source	Target
En \rightarrow De	72	1550	1444	3.2	6.1
En \rightarrow De ²	38	632	572	3.5	6.8
En \rightarrow Es	65	1739	1815	5.6	6.3
En \rightarrow Fr	81	2226	2444	7.0	7.2
En \rightarrow It	28	783	783	3.9	4.2
En \rightarrow Sl	20	257	210	0.6	1.1

systems outperform multilingual systems in all cases. However, the gap between bilingual and multilingual is almost one BLEU point smaller when translating into Slovenian, since the Slovenian target language of all four source languages is combined into a single system.

Table 15: BLEU scores achieved by bilingual and multilingual Slovenian-pivot systems.

		VL		IWSLT			
		dev	test	dev2012	test2012	test2013	test2014
Bilingual	Sl \rightarrow En	30.5	26.4	32.1	31.4	36.0	34.3
	En \rightarrow Sl	27.8	22.9	27.8	25.8	29.4	27.5
Multilingual	Sl \rightarrow En	26.8	19.6	30.9	29.9	33.9	32.7
	En \rightarrow Sl	26.3	19.7	26.3	25.0	27.5	25.4

Table 16 reports BLEU score provided by translation system from English into German, Spanish, French, Italian and Slovenian. Following the same trend as in Table 15, our best bilingual systems reported in deliverable D3.5 outperform the English-source multilingual system. However, the gap between bilingual and multilingual systems in the case of translating into Spanish and Italian ranges from 1.8 to 3.2 BLEU points, followed by German and Slovenian from 3.0 to 6.3 BLEU points, and finally French with a significant gap over 10 BLEU points. We can also compare the BLEU score of the English into Slovenian translation for both multilingual systems, slightly outperforming the Slovenian-pivot system (19.7) the translation accuracy of the English-source system (19.1).

The multilingual systems here reported are our first effort to deploy systems that can translate from N to M languages, even though the results are promising and further analysis is required to gain experience and understand the disparity of BLEU scores obtained across languages. It is also worth mentioning that the multilingual English-source system trained with over 300 millions pairs is the largest MT system we have ever trained and took to the limit our computing resources. In this regard, more computational efficient models need to be devised and assessed to take advantage of even larger corpora in multilingual setups.

Table 16: BLEU scores achieved by bilingual and multilingual English-source systems.

En→	dev					test				
	De	Es	Fr	It	Sl	De	Es	Fr	It	Sl
Bilingual	30.3	38.0	35.1	28.9	27.8	45.7	34.6	41.1	29.8	22.9
Multilingual	27.3	36.2	23.7	26.5	22.7	39.4	32.2	28.7	26.6	19.1

7.4 Cross-lingual text-to-speech dubbing

The more recent results from X5gon, obtained by application of the very latest developments in ASR/MT to large OER repositories, have shown that we have now reached the point at which raw subtitles are often good enough for direct publication, particularly in the source language [18, 5]. Moreover, as shown in the preceding sections on streaming ASR and simultaneous MT, current research in ASR/MT is also showing that advanced systems no longer need prerecorded audio (*offline* setup), as they can now work with no significant degradation under the so-called *streaming* setup; that is, subject to the constraint that output must be delivered in nearly *real time*, only within a short delay after the incoming audio stream [34]. To us, all of this recent progress will soon result in a rapid increase of (raw) multilingual subtitles of publishable quality for large repositories of educational videos, and also *live* lectures, either online or not, delivered under reasonable acoustic conditions.

Assuming a progressive reduction of the cost to produce publishable subtitles in diverse target languages, it is natural to also consider the use of text-to-speech (TTS) tools to efficiently dub the lecturer’s speech in target languages she/he might not even speak. In fact, as with subtitles, synthesized speech has been used for many years to make content accessible to people with disabilities [35]. Other areas for which TTS tools have provided support include second language learning [36], reading difficulties [37] and virtual humans [38]. Although these tools have been available and used for many years, it has not been until very recently that a plethora of contributions based on modern AI tools have dramatically improved and extended TTS capabilities. Indeed, the naturalness of the speech generated by state-of-the-art TTS systems is now known to rival that of human speech [39]. Also, the most advanced TTS systems are capable of generating speech for multiple speakers and languages, even in the usual case in which speakers can only provide training data in just a few languages, thus enabling cross-lingual voice cloning in all target languages of interest [40]. To us, this particular feature is especially interesting to bridge language barriers at universities, as it opens the door to produce multilingual educational videos at scale with both publishable subtitles and cloned lecturer speech.

In this section, we report the experience gained on (cross-lingual) voice cloning by the UPV in recent years, and especially in Y3 within the context of X5gon’s Task 5.2. It builds on past and more recent (X5gon) work on using modern AI tools to produce multilingual subtitles and synthesized speech for the UPV’s main repository of educational videos, *poliMedia*, also known as *MediaUPV* in its extended version including all kind of educational videos produced at the UPV [41, 4]. Our first, pioneering tests using deep neural networks (DNNs) for Spanish TTS in MediaUPV were carried out by the end of the European project transLectures [42]. Albeit with some delay with respect to ASR and MT, at that time it was clear to us that TTS technology was on the brink of a breakthrough on both performance and capabilities. Thus, in order to properly assess what TTS progress can do for voice cloning at the UPV, two main actions were taken. On the one hand, a call for participation to the UPV’s academic staff was made so as to collect *clean* lecturer speech data, and later survey their opinions and suggestions on the potential application of TTS at the UPV. At this point it is worth noting that acquiring such a database of lecturer speech data was also seen as crucial in learning about patterns of language proficiency among UPV lecturers with good predisposition to use TTS. On the other hand, we began to monitor TTS progress, especially as regards to systems capable of dealing

with multiple speakers and languages. After the acquisition of the database during the first half of X5gon, a multilingual and multi-speaker TTS system was built from current state-of-the-art TTS technology adapted to the UPV case. This system was then used to voice-clone (part of) MediaUPV and survey what UPV lecturers participating in our study think about present TTS technology. For the survey, participants listened to human and synthetic voice, for their own and others' videos in MediaUPV, also including cross-lingual cloned voice. Although the survey originated many questions and thoughts from lecturers, the general view is that TTS technology is not only mature enough for its application at the UPV, but also needed as soon as possible, especially to bridge language barriers for foreign students.

In what follows, we begin with a description of our lecturer speech database for TTS, its acquisition protocol and basic statistics (Section 7.4.1). Section 7.4.2 follows with a review of our production pipeline of subtitles and cloned voice, particularly in regards to TTS technology, its state-of-the-art and the way it was adapted to train a multilingual and multi-speaker TTS system from our lecturer speech database. Section 7.4.3 is devoted to the evaluation of this TTS system, the protocol and support platform we used to acquire the UPV lecturers' opinion on it, and the results obtained. Finally, some concluding remarks are given in Section 7.4.4.

7.4.1 The DeX-TTS dataset

Although poliMedia (as part of MediaUPV) was already described in the X5gon proposal, the interested reader is referred to Section B for an updated description, as of June 2020, also written from a linguistic perspective and paying attention to the way publishable multilingual subtitles have been produced cost-effectively. Concerning the acquisition of our lecturer speech database for TTS, a call for participation to the UPV's academic staff was made under the DeX plan to collect *clean* lecturer speech data during the first half of X5gon approximately, which was answered by a total of 98 participants. To this end, a number of sentences in Spanish, Catalan and English were first drawn from various sources (mainly newspapers, MOOCs and Wikipedia) and then reviewed for readability. Similarly to poliMedias, speech recordings were made under the same acoustic conditions at poliMedia studios, during two 90-minute sessions per participant. Participants were asked to record a minimum of 300 randomly drawn sentences in either one or two languages (with a minimum of 150 in each). In reality though, they were encouraged to record as many sentences as possible within the time available, not only in their mother tongue (typically Spanish or Catalan), but also in the other two languages under consideration, even if low-proficient (which is often the case in English); indeed, they were allowed to skip sentences when unsure about their correct pronunciation. As shown in Table 17, the net effect of this encouragement was more participants contributing in multiple languages rather than just one, which is different from what happens with poliMedias themselves (see Table 29), though good for our purposes.

Table 17: Participants contributing to clean speech data collection in Spanish (es), Catalan (ca), English (en), bilingual combinations (es-ca, es-en, ca-en) and the trilingual case.

	Monolingual			Bilingual			Trilingual	Total
	es	ca	en	es-ca	es-en	ca-en	es-ca-en	
Participants	36	1	4	16	22	3	16	98
Total	41			41			16	98

Table 18 shows the number of sentences and duration in hours collected in our *DeX Text-To-Speech (Dex-TTS)* dataset of clean lecturer speech data. In total, it comprises 59 hours of clean speech data from 47K sentences uttered by 98 participants. Looking at it row by row, it can be seen that Spanish,



Catalan and English account for around 61%, 15% and 24% of the data (both in terms of sentences and recorded speech), respectively. By columns, we can observe that most of the data comes from multilingual acquisitions, either bilingual (42%) or trilingual (23%), meaning that only some 35% of the data corresponds to monolingual participants.

Table 18: Number of sentences and duration in hours of the clean speech data collected in Spanish (es), Catalan (ca), English (en), bilingual combinations and the trilingual case.

		Monolingual			Bilingual			Trilingual	Total	%
		es	ca	en	es-ca	es-en	ca-en	es-ca-en		
No. of sentences ($\times 1000$)	es	14.6	-	-	4.1	6.7	-	3.5	28.9	61
	ca	-	0.3	-	2.7	-	0.5	3.8	7.3	15
	en	-	-	1.0	-	5.5	0.6	4.0	11.1	24
Total		15.9			20.1			11.3	47.3	-
%		34			42			24	-	100
Duration in hours	es	19.2	-	-	5.4	8.0	-	3.7	36.3	62
	ca	-	0.4	-	3.4	-	0.6	4.1	8.5	14
	en	-	-	1.3	-	6.9	0.7	5.1	14.0	24
Total		20.9			25.0			12.9	58.8	-
%		36			42			22	-	100

The DeX-TTS dataset is undoubtedly a very valuable resource to test modern TTS technology at the UPV and also an example that can be easily replicated in other universities. On the one hand, TTS technology does not require vast amounts of manually transcribed speech data, as ASR does, but simply a relatively small corpus of clean speech. Indeed, our corpus is similar in size to those commonly used in TTS research (cf. [39] and [43]). On the other hand, being produced at the UPV by its academic staff, the DeX-TTS dataset is an optimal resource to explore how a UPV lecturer's speech can be best cloned, not only in her/his mother tongue, but also in other languages she/he might not even speak. In this regard, our corpus can be considered a good example of linguistic diversity at a higher education institution, where the dominant official language (Spanish) coexists with a minority yet official language (Catalan) and English. As a result, the DeX-TTS dataset is rich in Spanish speech data but no so rich in Catalan and (non-native) English speech.

7.4.2 Cross-lingual voice cloning at the UPV

As described at the beginning of Section 7.4, our work on (cross-lingual) voice cloning at the UPV relies on modern AI tools to produce cost-effective multilingual subtitles and synthesized speech for poliMedias. This is clearly illustrated by the production pipeline diagram shown in Figure 22. The process begins with a new poliMedia uploaded to MediaUPV, including its speaker (lecturer) and (source) language IDs. The first pipeline step (ASR) consists in automatically transcribing the new poliMedia to produce raw source subtitles, which can be optionally reviewed (post-edited) if convenient. In the second step (MT), source subtitles (transcriptions) are machine-translated into a number of target languages (e.g. into Catalan and English if the source language was Spanish). As with transcriptions, target subtitles (translations) can also be post-edited if convenient. TTS comes as the third and final pipeline step; in it, the speaker is automatically voice-cloned (dubbed) for each target language from the corresponding target subtitles. Note that each of the three pipeline steps requires specific models that need to be trained in advance from appropriate training data. The reader is referred to [41] for more details on the first two steps of the production pipeline. In what follows, our focus is on the TTS step, for which we assume (reviewed) translations to be available in each target language of interest.

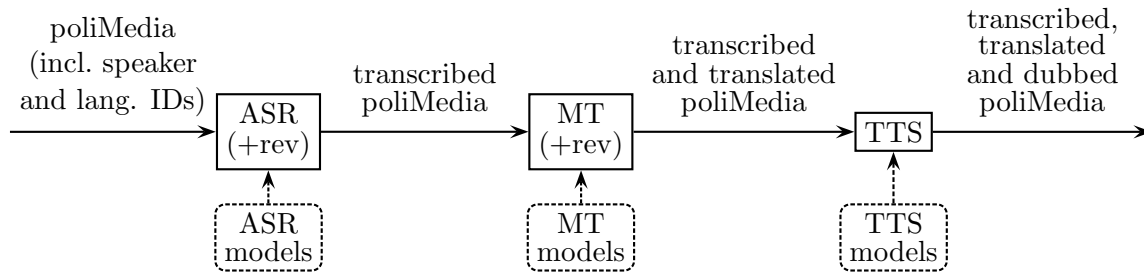


Figure 22: Production pipeline of transcribed, translated and dubbed poliMedias.

Until a few years ago, conventional TTS systems consisted of diverse, handcrafted components requiring highly specialized expert knowledge of both acoustics and linguistics. Moreover, they were normally restricted to a single speaker and language, making them impractical for massive voice cloning even in just a single language. However, driven by the deep learning revolution and an increased interest among big technology companies, the field of TTS has recently seen large improvements in quality, flexibility and capabilities. In brief, conventional TTS approaches have been surpassed by *end-to-end neural network architectures* [44, 39, 43] and *neural vocoders* [45, 46]. In particular, Google’s *Tacotron2* has become the *de facto* standard architecture for end-to-end TTS [39]. Compared to previous TTS technology, end-to-end TTS does not require highly specialized expert knowledge, achieves higher degrees of speech naturalness [39], and can be easily extended to deal with the general multilingual and multi-speaker setting [40]. As discussed at the beginning of Section 7.4, this generality is a key feature of the new end-to-end neural architectures, as it opens the door to massive machine dubbing of educational videos, even in target languages of which the speaker has little or no command. With this idea in mind, an extension of Tacotron-2 for multiple speakers and languages was developed after completing the DeX-TTS dataset, which is referred to below as *Tacotron2-UPV*. Its basic architecture is depicted in Figure 23.

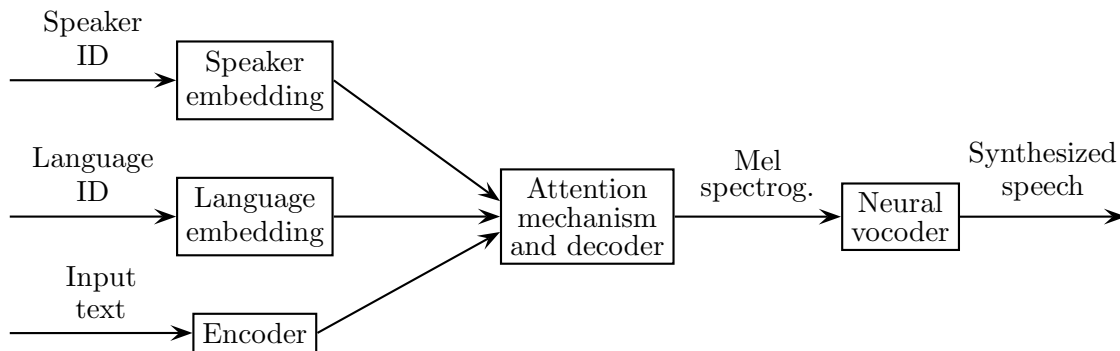


Figure 23: Basic Tacotron2-UPV architecture.

As can be seen in Figure 23, the Tacotron2-UPV system consists of five main components, each of them in turn comprising a number of neural-network components (omitted for simplicity). Given a speaker ID, language ID and the input text (target subtitles) to synthesize, they are first processed by, respectively, the *speaker embedding*, *language embedding* and *encoder* components. Their output is then combined into an internal, compact data representation which (hopefully) captures everything that is relevant in the raw input to synthesize highly natural speech. This is actually done next, in two consecutive steps. In the first step, an *attention mechanism* and a *decoder* deal with the conversion of the encoded input into a (*mel-scale*) *spectrogram* of the target audio waveform, which is basically a compact representation still retaining sufficient intelligibility and prosody information. Finally, in the

second step, a *neural vocoder* produces the desired synthesized speech from its spectrogram.

At this point it is worth noting that the Tacotron2-UPV system was conceived and developed after completing the DeX-TTS dataset, independently of Google's own multilingual and multi-speaker extension to Tacotron2 reported in [40]. There are not many differences, however, between the two systems. Indeed, the most salient difference is in the way multiple languages are taken account of. In Tacotron2-UPV, separate grapheme embeddings per language are used to capture language-dependent particularities at shallow system layers (in the language embedding component), thus facilitating deeper layers (in the attention mechanism and decoder component) to focus more on language-independent patterns of the human voice. In contrast to this, a common set of grapheme/phoneme embeddings for all languages is used by [40].

As with end-to-end TTS models in general, Tacotron2-UPV can be trained with minimum human intervention (and expert knowledge) from an appropriate collection of <text, audio> pairs. In this regard and as noted above, the DeX-TTS dataset is a very valuable resource as it was acquired with this goal in mind. However, also as noted above, it is rich in Spanish but no so rich in Catalan and English, and thus a TTS system trained only from it will certainly be biased towards Spanish. This is not likely to be an issue for Catalan due to its high similarity to Spanish. However, it is certainly an issue for English, not only because of its comparatively lower degree of similarity, but also due to the limited level of fluency in the non-native English speech recorded. To compensate for this lack of (fluent) English speech data, we also included (part of) the *VCTK* corpus of multi-speaker native English speech for TTS [47].

The actual training of the Tacotron2-UPV system was carried out after applying a few common preprocessing steps for TTS data. In particular, the DeX-TTS dataset was preprocessed by first trimming leading and trailing silences, and then applying certain basic audio filters to reduce noise and loudness variability among recordings. All Tacotron2-UPV components but the neural vocoder were jointly trained using an extended version of a publicly available implementation of the basic Tacotron2 [48]. Similarly, the neural vocoder was trained using an open-source implementation of WaveRNN [46] by [49]. In this way, a complete, fully-trained Tacotron2-UPV system was built to enrich any poliMedia with machine-dubbed audio tracks in its target languages. In this regard, it is worth mentioning that, for the synthesized speech to be (more or less) in synchrony with the video image, machine dubbing is done at the sentence level and aligned in time with source sentences. It also must be noted that, although Tacotron2-UPV was developed thinking primarily about contributors to the DeX-TTS dataset, it can be applied to poliMedias by other authors as well by simply choosing appropriate target speakers.

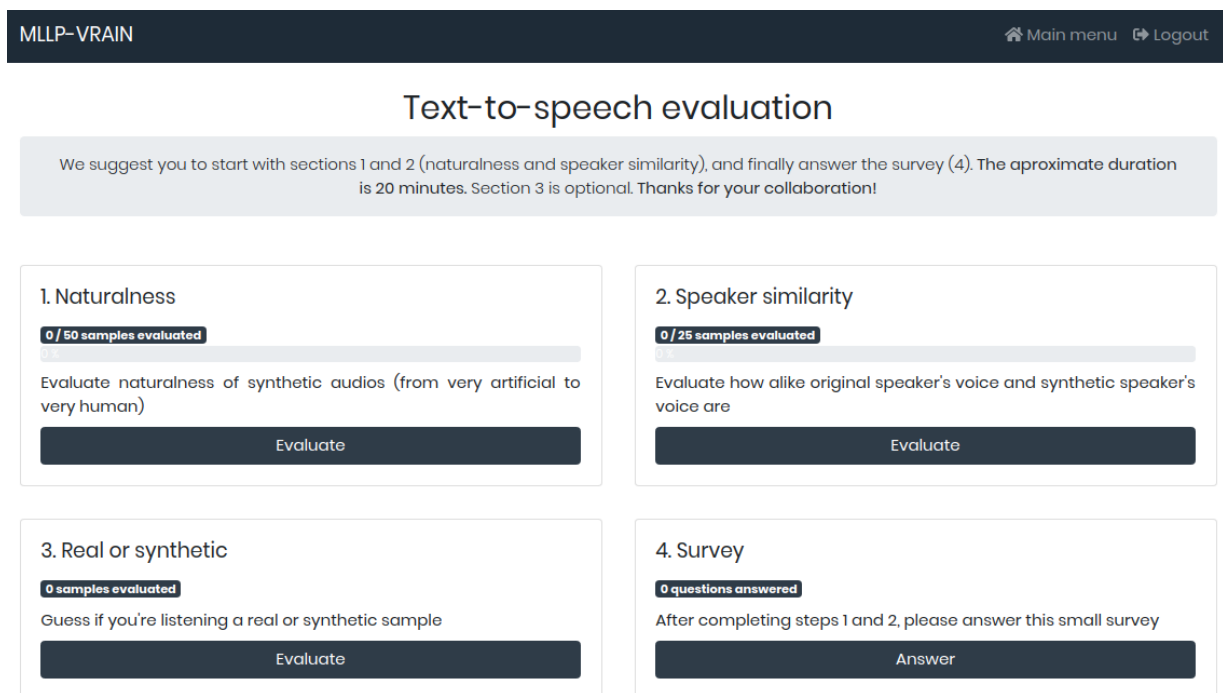
7.4.3 Evaluation

Evaluation of machine learning progress by machine learners is generally driven by widely accepted, *objective* (well-defined) metrics that can be automatically computed by comparing system output and *ground truth* on a set of data samples not used for system training (*test set*). Being able to compute objective metrics in a fully automatic way is seen as a key factor to speed up progress, since not only can researchers thus compare their achievements easily and objectively, but also production of new, improved systems is accelerated by simply running a fully-automated training and testing loop. A good example of this is the WER metric, which has successfully driven the ASR field for decades [50]. Analogously, the BLEU accuracy measure [51] and the WER-inspired *Translation Edit Rate (TER)* metric [52] have played a similar role in MT. Needless to say, most important of all for objective metrics is to be highly-correlated with human judgement.

In contrast to ASR and MT, no objective metrics have gained wide acceptance in TTS and, indeed, most recent work is assessed only by means of *subjective* evaluations [44, 39, 43]. Generally speaking, (listening-type) subjective evaluations boil down to human participants listening to (real

and synthetic) speech utterances and giving their feedback on the speech quality, either globally or in terms of individual factors. More precisely, the ITU-T Recommendation P.85 [53] is at the basis of most testing methods used for evaluating the subjective quality of synthetic speech. In it, the recommended testing method consists in asking subjects to express their opinion using one or more five-point opinion (Likert) scales. In addition to the overall quality scale, other scales can be considered for measuring listening effort, voice pleasantness, etc. However, by far the preferred way to test and compare current TTS systems is in terms of overall quality only, and on the basis of a *mean opinion score (MOS)* with a 95% confidence interval [44, 39, 43].

To assess the Tacotron2-UPV system described in Section 7.4.2, a call for participation was made to the 98 lecturers contributing to the DeX-TTS dataset (Section 7.4.1), which was answered by nearly half of them (47). The evaluation procedure was designed around a *test set* of 8820 speech samples synthesized by Tacotron2-UPV. They correspond to 98 lecturers, times 3 languages per lecturer, times 30 sentences for each lecturer-language pair, with sentences randomly picked from poliMedia subtitles not used for Tacotron2-UPV training. Note that many test samples were produced by cross-lingual voice cloning since nearly half (42%) of all lecturer-language pairs were not covered by training data in the DeX-TTS dataset (see Table 17). With this test set at hand, participants were asked to register at a web platform for them to proceed with the evaluation from a user home page (Figure 24).



MLLP-VRAIN [Main menu](#) [Logout](#)

Text-to-speech evaluation

We suggest you to start with sections 1 and 2 (naturalness and speaker similarity), and finally answer the survey (4). The approximate duration is 20 minutes. Section 3 is optional. Thanks for your collaboration!

1. Naturalness

0 / 50 samples evaluated

Evaluate naturalness of synthetic audios (from very artificial to very human)

Evaluate

2. Speaker similarity

0 / 25 samples evaluated

Evaluate how alike original speaker's voice and synthetic speaker's voice are

Evaluate

3. Real or synthetic

0 samples evaluated

Guess if you're listening a real or synthetic sample

Evaluate

4. Survey

0 questions answered

After completing steps 1 and 2, please answer this small survey

Answer

Figure 24: Home page of the evaluation platform.

As shown in Figure 24, the evaluation procedure consisted of four parts: *1. Naturalness*, *2. Speaker similarity*, *3. Real or synthetic* and *4. Survey*. It was suggested to start with parts one and two, then optionally move to part three, and finally answer the survey in part four. With the help of a brief progress indicator in each part, participants were allowed to stop and resume the procedure as they wished. In what follows, procedural details and evaluation results are provided for each part separately.

Naturalness

Naturalness refers to overall speech quality, that is, the main criterion by which current TTS systems are tested and compared. Using a five-point (star) opinion scale, participants were asked to rate the naturalness of a minimum of 50 samples randomly drawn from the test set (Figure 25). For validation purposes, truly natural (human) speech recordings were also included as control samples among synthetic ones, at random with a ratio of one human recording per six evaluated samples.

Evaluation progress:

17 / 50 evaluated samples
34 %

Text:

You can do that for the border cases and then you get something like this.

Audio:

0:00 / 0:03

Naturalness:

★★★★☆ 4.0

☐ The audio and the text do not match or the audio presents a significant anomaly.

Confirm and continue

Figure 25: Naturalness evaluation interface.

Table 19 shows, for each language, the naturalness MOS with 95% confidence intervals for both synthetic and control samples, as well as the number of evaluated samples. The *seen* and *unseen* columns refer to synthetic samples from lecturer-language pairs used and not used, respectively, for Tacotron2-UPV training.

Table 19: Naturalness MOS with 95% confidence intervals per language, including cross-lingual cloning (synthetic samples from lecturer-language pairs *unseen* in training).

Language	Naturalness MOS			Control samples	Evaluated samples
	Seen	Unseen	Total		
Spanish	4.1 ± 0.1	3.9 ± 0.3	4.1 ± 0.1	4.5 ± 0.2	533
Catalan	4.2 ± 0.1	4.0 ± 0.1	4.1 ± 0.1	4.8 ± 0.1	551
English	3.6 ± 0.2	3.6 ± 0.1	3.6 ± 0.1	4.3 ± 0.2	594

From the results in Table 19, it can be observed that the naturalness MOS on the synthetic speech produced by Tacotron2-UPV is in general fairly good though, as expected, not as good as human speech. In particular, the naturalness of synthetic Spanish and Catalan was judged to be at the very same high rate of 4.1, slightly but significantly below that of human Spanish (4.5) and Catalan (4.8). Similarly, the naturalness of synthetic English was rated at 3.6, again slightly but significantly below that of human speech (4.3). These comparatively lower rates for (synthetic and human) English are certainly due to the non-nativeness nature of the English recordings in the DeX-TTS dataset, from which we get, not surprisingly, a (realistic) non-native bias for English in Tacotron2-UPV. In any case, summarizing, a main conclusion from Table 19 is that Tacotron2-UPV produces highly natural synthetic speech, not far from human speech. Moreover, by comparing the seen and unseen rates for each language, we see that, in general, synthetic speech naturalness does not depend significantly on

which specific lecturer-language pairs were covered in the training data. In other words, Tacotron2-UPV has effectively learned to transfer (clone) lecturer voices from source languages (e.g. mother tongue) to target languages they might not even speak.

Speaker similarity

Although naturalness is without question the main criterion to judge synthetic speech goodness, it falls short in measuring how similar original (human) and cloned (synthetic) voices actually are. This is particularly relevant for cross-lingual voice cloning since, as pointed out above, it seems that Tacotron2-UPV is capable of cloning voice for unseen lecturer-language pairs almost as well as for seen ones. Needless to say, as this is a feature only available to the most advanced TTS systems, it deserves empirical confirmation. To this end, the second part of the evaluation procedure consisted in rating, on a five-star opinion scale, the speaker similarity between a test sample picked at random, and a training sample also picked at random from the same speaker but not necessarily from the same language. Participants were asked to do this for a minimum of 25 test samples. Table 20 shows the speaker similarity MOS with 95% confidence intervals for the seen and unseen cases separately, and the number of evaluated samples.

Table 20: Speaker similarity MOS with 95% confidence intervals per language, for test samples produced from seen and unseen lecturer-language pairs of training data.

Language	Speaker similarity MOS		Evaluated samples
	Seen	Unseen	
Spanish	4.2 ± 0.1	4.0 ± 0.5	324
Catalan	4.1 ± 0.2	4.0 ± 0.2	284
English	3.7 ± 0.2	3.4 ± 0.2	299

From the results in Table 20, we can confirm that cross-lingual voice cloning by Tacotron2-UPV works almost as well as conventional voice cloning from seen lecturer-language pairs. Although minor (not significant) yet consistent MOS differences show a slight preference for cloned voice in the seen case, to us this is rather a confirmation that current TTS technology can be safely used for cross-lingual machine dubbing.

Real or synthetic

As an extra check to validate MOS results on naturalness and speaker similarity, participants were also invited to optionally run a sort of Turing test to try to guess whether a given speech sample is real (human) or synthetic. This was done in the third part of the evaluation procedure, from speech samples picked at random with a ratio of two synthetic samples per each real one. Table 21 shows accuracy (success rate) results per language for real, synthetic and all (total) samples, along with the number of evaluated samples.

Table 21: Participant accuracy on the *real or synthetic* test.

Language	Accuracy (%)			Evaluated samples
	Real	Synthetic	Total	
Spanish	79	52	63	121
Catalan	72	59	63	90
English	66	68	67	101
All	73	60	64	312



Although the number of evaluated samples is modest, from the results in the total column of Table 21, we see that participants were not more accurate than just classifying all samples as synthetic (which would result in a total accuracy rate of 67%). Of course, participants were not aware of the prior probability (ratio) by which two of every three samples were synthetic. Instead, it seems that they tried to guess the true origin of each given sample by implicitly assuming that the two possible origins were equally probable. Interestingly, in doing so they were more accurate in spotting real samples than synthetic ones in the languages they know best. All in all, these results again confirm that the quality of the speech synthesized by Tacotron2-UPV is really close to human speech.

Questionnaire and comments

The fourth and final part of the evaluation procedure consisted of just two Yes or No control questions on the acceptance of TTS technology, each accompanied by a box for free-text comments and suggestions. Table 22 shows these two control questions and the Yes or No votes received.

Table 22: Final questions and answers on the acceptance of TTS technology.

<i>Questions:</i>	<i>Yes</i>	<i>No</i>
Do you think that the shown automatic dubbing technology can be useful to improve accessibility and engagement in online educational materials?	47	0
Would you accept your educational materials to be automatically dubbed in different languages using this technology?	46	1

As shown in Table 22, all participants think that machine dubbing is useful to improve accessibility and engagement in online educational materials. Also, almost all of them would accept their educational materials to be automatically dubbed in different languages using Tacotron2-UPV.

Apart from the Yes or No feedback, each question originated many comments by participants. On the one hand, we received sixteen comments to the first question: four of them pointed out that there is still room for improvement in pronunciation, nine others were just very positive feedback on the speech synthesis quality and, finally, three comments suggested extending our work to *full* machine translation of poliMedias including slides. On the other hand, thirteen comments were made to the second question: seven of them were to encourage us to deploy TTS technology into production without delay, while the six other comments just requested that lecturers be allowed to review and approve their machine-dubbed materials prior publication. Summarizing, the general view of our study is that TTS technology is not only mature enough for its application at the UPV, but also needed as soon as possible.

7.4.4 Concluding remarks

This work has reported the experience gained on the use of TTS technology at the UPV in recent years, and especially in Y3 within the context of X5gon's Task 5.2. We have first focused on the main data resource needed to build an in-house, repository-adapted (cross-lingual) TTS system: our lecturer speech database for TTS, its acquisition protocol and basic statistics. This has been followed by a review of our production pipeline of subtitles and cloned voice, particularly in regards to TTS technology, its state-of-the-art and the way it was adapted to train a multilingual and multi-speaker TTS system from our lecturer speech database. Finally, an extensive, subjective evaluation of this TTS system has been reported, including the protocol and support platform we used to acquire the UPV lecturers' opinion on it, and the results obtained. Summarizing, these results show that TTS

technology is mature enough for massive machine dubbing of educational videos, even in the cross-lingual case. To us, the door has been opened to producing multilingual educational videos at scale and low cost with both publishable subtitles and cloned lecturer speech of high quality.

8 Conclusions

In this deliverable, we have reported the work done in Task 5.2 from M25 to M36, and also during X5gon's four-month extension to complete the user studies led by UCL (M37–M40). Thus, for the purpose of this report, Y3 refers to the period from September 2019 (M25) to December 2020 (M40).

An important starting point for this report is M24, in which JSI finalised the integration of the three major (planned) project components (X5oerfeed, X5analytics and X5recommend) into the *X5gon platform* under different names (described in deliverable D2.2). For coherence with D5.1 and D5.2, however, we have still used the term X5oerfeed to refer to OER pipeline processing services, particularly (video) automatic transcription and translation services developed by UPV, as well as the term X5recommend for the X5gon recommender engine. The X5analytics component, whose development has been more difficult than anticipated, was from M25 on (Y3) integrated into the X5gon platform through an API allowing access to multiple analytics, models and tools. In contrast to X5oerfeed and X5recommend, which have been extensively piloted along the whole project, X5analytics has been mainly assessed in Y3 as part of the user studies carried out by UCL.

Following the structure of this report, it is worth highlighting the following results and conclusions:

1. *VideoLectures.Net*. Being the first and largest official pilot in X5gon, Videolectures.NET has been a primary focus of interest to WP5 during the whole project. In brief, some relevant results and conclusions worth to emphasize from Videolectures.NET piloting are the following:
 - *Transcription error*. The quality of automatic transcriptions for Videolectures.NET videos, most in English and Slovene, has been considered a major objective since the very beginning of X5gon. In this regard, we have managed to reduce the transcription error from 19.6% (M0) to 14.8% (M40) in English and, furthermore, from 32.5% (M0) to 15.3% (M40) in Slovene. To us, crossing into the “safe” area of error rates below 20% is clearly a major breakthrough for Videolectures.NET and X5gon. Indeed, for the reader to get an idea of how difficult this AI challenge is, Google Cloud Speech-To-Text on Videolectures.NET videos only attains error rates of 28.6% and 50.0% in, respectively, English and Slovene.
 - *Translation accuracy*. Transcription error rates below 20% not only mean publishable subtitles at scale and low cost for Videolectures.NET, but also enabling a number of other applications that can be derived from accurate subtitles, and especially their translation into other languages. This is, of course, another major objective since the very beginning of X5gon. Although many advanced MT systems have been developed using English as a pivot language, in most cases translation accuracy was more or less on par with that of Google Translate. A notable exception, however, occurs with English and Slovene: when compared to Google Translate, our systems achieved relative improvements of 76% for Slovene→English and 39% for English→Slovene. Needless to say, opening up OER in minority languages such as Slovene is not a commercial priority to mainstream providers of language technology.
 - *Recommender*. Although the X5gon connect service was only deployed in VideoLectures.Net and poliMedia, the X5gon network of OER repositories has been notably enlarged by many worldwide OER repositories being indexed from the X5gon platform. This has really helped to pilot most X's of X5gon from nearly 1.3M records of user transitions via the



X5recommend plugin. In particular, from the VideoLectures.Net perspective, it was found that nearly one of every four transitions were to OER from other sites.

2. *poliMedia*. The second official pilot in X5gon has clearly shown the great value of fully deploying X5gon's tools and services into a higher education institution. To a large extent, the X5gon experience at the UPV parallels that of VideoLectures.Net:

- *Transcription error*. As with Videolectures.NET, we have focused on two *source* languages: Spanish, the dominant language in poliMedia, and English. In contrast to Videolectures.NET, from the very beginning of X5gon, we already had automatic systems achieving fairly good transcription error rates (11.7% in Spanish and 22% in English). Nevertheless, we also managed to greatly reduce these rates in M40, down to 8.3% in Spanish and 12.0% in English. As before, to get a sense of how challenging this is, Google Cloud Speech-To-Text on poliMedia only delivers 19.9% for Spanish, and 36.1% and 13.3% for English in, respectively, standard and enhanced mode. Summarizing, although automatic transcription of poliMedia videos was more or less solved in M0, we managed to get it largely solved in M40.
- *Translation accuracy*. In the poliMedia case, our focus was on Spanish→English MT though, as with most language pairs tried (with English involved), results were more or less on par with Google Translate. In connection to this, it is worth to mention that automatic transcription of Spanish poliMedias (at 8.3% of WER), followed by their automatic translation into English (at 34.1% of BLEU), is the best case we have piloted for fully automatic transcription and translation of OER. To us, it clearly shows that language technology is mature enough to provide functional (publishable) multilingualism to OER across Europe (and worldwide).
- *Recommender*. By deploying the X5gon connect service into poliMedia as early as Y1, UPV became the first higher education institution providing cross-lingual, cross-modal and cross-site X5gon recommendations to its students. From Y1 on, X5gon and UPV's own (internal) recommendations have been mixed at random in equal proportions and offered to students accessing the UPV media portal. Similarly to VideoLectures.Net, it was found that one of every three user clicks were on X5gon recommendations. Taking into account that UPV students are typically taught to follow instructors' learning paths within poliMedia, this result is not bad at all. On the contrary, it clearly shows that a large network of collaborative multilingual OER sites has a great potential in boosting educational opportunities for all. In our analysis of the X5gon recommendations followed by UPV students, we found that limited yet significant part of them were effectively cross-site (18%), cross-lingual (26%) and cross-modal (9%). Again, we tend to think this result is not bad at all, especially given that UPV students are usually asked to follow instructors' learning paths within poliMedia.

3. *virtUOS*. UOS, the third official pilot in X5gon and a higher education institution like UPV, has showcased what we can expect from many higher institutions: the X5gon initiative is absolutely great but, for it to work in practice, lecturers need effective recommendation and search tools delivering high-quality OER content they may recommend to their students. Due to this, UOS proposed an "OER Recommender for Lecturers" pilot in Y1, which in Y2 was updated to its "X5gon Discovery Pilot", and then run in Y2 and extended to Y3. In brief, UOS found that X5gon recommendation and search tools have been greatly improved along the project, though there is still room for improvement in terms of indexed OER (quantity and) quality.

4. *Other pilots*. Although WP5 focus was mainly on our three official pilots, X5gon partners from higher institutions not involved as official pilot providers, i.e. UCL and Nantes, also modestly



helped to pilot X5gon tools and services from their own and external OER. In this regard, it is worth mentioning the great effort made by these partners, particularly Nantes, to link X5gon to the large community of Moodle, though we agreed to frame and report this effort within the work package devoted to studies in the wild (WP6). Apart from these other pilots, we would like to emphasise that X5gon ideas and tools have attracted many educational stakeholders from diverse origins and interests. A good example of this is the Kobi app experience run in Y3, by which our automatic transcription tool for Slovene was applied to help children with reading difficulties in improving their reading skills.

5. *User studies.* The first main subtask of Task 5.2 was to pilot advanced analytics and social context meetings. As indicated above, to this this end UCL has conducted a number of user studies using the *X5Learn* platform also developed at UCL. These studies were intended to test a number of innovative and advanced interaction features of X5Learn: search engine, content flowbar, views, playlists and note taking. However, due to unexpected late interruptions (mainly resulting from the COVID-19 pandemic), only the content flowbar and playlists features were effectively tested by means of separate pilots:

- *Content Flow Bar (CFB) Pilot Study.* The CFB was developed to facilitate video browsing by providing semantic “snippets” related to content popping up on screen from the navigation flow bar. Broadly speaking, the study consisted in asking participants to look for relevant OER on a certain topic and then interviewing them about the CFB usability. Although it was carried out while still developing the tool, participants were very positive about the CFB.
- *Playlist Pilot Study.* From a learning analytics perspective, the playlists feature in X5Learn was without doubt the most relevant instrument for piloting. In its study of playlists, UCL performed a thorough analysis on how best X5Learn, and its back-end tools and services from the X5gon platform, can be used to construct optimal learning paths for users. After an initial pre-study with university lecturers, the actual study was conducted with high-school maths instructors asked to evaluate the tool for math teachers’ training. Overall, it was positively evaluated and, indeed, participants ended up thinking about how they could use it to construct playlists for their own use.

6. *Advanced cross-lingual and cross-modal features.* The second main subtask of Task 5.2 was to pilot advanced cross-lingual and cross-modal features. This subtask, led by UPV, has been divided into four research lines: streaming automatic speech recognition (ASR), simultaneous machine translation (MT), multilingual MT, and cross-lingual text-to-speech (TTS) dubbing. Although each of these lines was considered of great interest on its own, the rationale to consider them all was to explore whether an accurate pipeline “streaming ASR → (multilingual) simultaneous MT → cross-lingual TTS dubbing” could be constructed for offline and *live speech-to-speech* translation of (*live*) audio streams. In X5gon, the kind of relevant applications we had in mind for such pipeline ranges from *fast* speech-to-speech translation of prerecorded (educational) videos (seen as a simple form of live audio sources) to multimodal annotation of OER (e.g. with learners’ comments), and live multilingual speech-to-speech translation on an educational platform (e.g. X5Learn), either for live lecturing or (live) user dialoguing (e.g. on particular OER). Summarizing, the main results and conclusions drawn from the research carried out along these lines are:

- *Streaming ASR.* In M25, we had already achieved low ASR error rates for Spanish, English and Slovene in our official pilots, and thus we were in a good position to try adapting our offline systems for these languages to work on live audio streams in real time (streaming

setup). To this end, a number of novel techniques were devised and tested on the pilots, showing that streaming adaptation is perfectly feasible at the expense of a minor relative error increase. Indeed, with the exception of an 11.7% relative error degradation on English poliMedias, we observed that the average relative error increase was only around 4%.

- *Simultaneous MT.* As with ASR, offline MT systems cannot be directly applied in the streaming setup since they require the full source sentence being available and no strict constraints on the response time. After studying very recent research proposals on simultaneous MT, two state-of-the-art techniques were chosen and empirically compared to our M24 and M30 offline MT systems for English↔Spanish on the well-known WMT task. In brief, the so-called *Multi-k framework* technique showed the best results, with only a minimal accuracy gap with respect to our offline systems. For instance, although offline systems achieved very good BLEU scores of 32.2 and 35.9 points (in M24 and M30, respectively) for Spanish→English, our (*Multi-k*) simultaneous MT system was not far behind with a BLEU score of 32.1.
- *Multilingual MT.* Following one of the most recent trends in neural-based MT, we explored the idea of building a single multilingual MT system capable of translating from N into M languages, thus replacing N-by-M MT separate systems otherwise needed if only (conventional) bilingual translation is considered. Apart from the obvious reduction in system building effort, this idea is also seen as a promising way to explore *transfer learning* in MT, that is, to transfer what is learnt from each separate language pair to each other. Both, effort reduction and transfer learning are really appealing in the X5gon context; indeed, being Slovene a low-resource language of high interest to us, we tend to think that transfer learning from rich-resource languages is the most promising way to boost MT accuracies for Slovene in the near future. Two Transformer-based multilingual systems were trained: one for Slovene↔{English, German, French and Spanish}, and another one for English→{German, French, Spanish, Italian and Slovene}. In short, multilingual systems are behind (M30) bilingual systems, though not much in many cases. For instance, the gap between them is only of 3.2 BLEU points for English→Slovene on VideoLectures.Net and, on WMT, 2.4 points for English→Spanish and virtually nothing (0.2 points) for English→Italian. It is also worth to mention that the Slovenian-pivot multilingual system enabled, for the first time in X5gon, Slovene↔{German, French and Spanish} direct MT (as opposed to indirect MT through English).
- *Cross-lingual TTS dubbing.* As discussed above in connection to the X5gon experience at the UPV, language technology is mature enough to provide fully automatic transcription and translation of OER with accurate output, as shown in X5gon for Spanish poliMedias translated into English. Thus, (cross-lingual) TTS dubbing is the final key piece needed to construct an accurate pipeline for (offline and *live*) *speech-to-speech* translation of (*live*) audio streams. In this regard, we have reported the experience gained on the use of TTS technology at the UPV in recent years, and especially in Y3 within the context of X5gon's Task 5.2. We have first focused on the main data resource needed to build an in-house, repository-adapted (cross-lingual) TTS system: our lecturer speech database for TTS, its acquisition protocol and basic statistics. This has been followed by a review of our production pipeline of subtitles and cloned voice, particularly in regards to TTS technology, its state-of-the-art and the way it was adapted to train a multilingual and multi-speaker TTS system from our lecturer speech database. Finally, an extensive, subjective evaluation of this TTS system has been reported, including the protocol and support platform we used to acquire the UPV lecturers' opinion on it, and the results obtained. Summarizing, these results show that TTS technology is mature enough for massive machine dubbing of

educational videos, even in the cross-lingual case. To us, the door has been opened to producing multilingual educational videos at scale and low cost with both publishable subtitles and cloned lecturer speech of high quality. This will be shortly followed by “game changing” streaming-adapted systems supporting *live* speech-to-speech translation in a number of educational applications.



A virtUOS: additional details

A.1 Sample JSON structures

Sample JSON structure for stored lecture and search results data:

```
{
  "lectures": [
    {
      "type": "lectures",
      "id": "1",
      "attributes": {
        "title": "Title in original language",
        "title_translated": "Title translated to en",
        "description": "Description in original language",
        "description_translated": "Description translated to en",
        "course_number": "3.210",
        "language": "en",
        "semester": "SS 2019",
        "faculty": "Faculty name",
        "results": ["*CONTAINS \rec_materials" FROM X5 API*]
      }
    }
  ]
}
```

Sample JSON structure for a survey data submits:

```
[
  {
    "uuid": "<generated UUID string>",
    "lectureId": 8,
    "resultId": 8,
    "localStorageKey": "x5pilot-l8-r8",
    "submitDate": "2020-01-22T15:04:43.327Z",
    "radioFit": 3,
    "radioSure": 3,
    "textComment": "<possible user comment>",
    "urlClickCount": 0,
    "isDuplicate": false,
    "modelType": ["tfidf", "wikifier"],
    "weight": [0.7063131928443909, 0.4518136978149414],
    "requestTime": 18341.925248,
    "lang": "en"
  }
]
```

A.2 Test set structure

A total of 22 courses from the summer semester 2019 were selected and the corresponding dates such as title, description and language were saved (see Figure 26 and Table 23). The structure of the test set is based on the course mix at UOS. Notable characteristics of the course information like included html-tags, included references or unusable description lengths were documented. This enables a subsequent analysis of possible error origins or sources of inaccuracies in OER recommendations.

Table 23: Test set data.

Courses sum	22
Search results per course	13.86
Possible survey results per course per participants	305

Table 24 shows the result structure according to the different model types that were used for the requests. For each model type, 110 results were stored and merged by double results ("materialId"). Corresponding data such as model type, weighting and query time were applied to the single results

id	faculty (translated)	title (translated to en if other lang)	has-description	lang-origin	tfidf results	wikifier results	doc2vec results	results (merged)	double results	specs/comment
0	Institute for Educational Science	Difference and diffusion of educational childhood and family childhood	yes	de	5	5	5	15	0	
1	Institute for Educational Science	Images of the Holocaust - Visual Media as Components of Contemporary Holocaust Learning	yes	de	5	5	5	10	5	
2	Institute for Educational Science	Digitisation in Early Childhood Education	yes	de	5	5	5	14	1	
3	Institute for Educational Science	Behaviour in school conflict situations	yes	de	5	5	5	13	2	long description
4	Anglistics	U.S.-American Literature and Culture from Modernism to Postmodernism	yes	en	5	5	5	14	1	
5	Neurobiology	Sensory Physiology	yes	en	5	5	5	15	0	short description
6	LE Cognitive Science	Action & Cognition II: Higher Cognitive Functions (Lecture)	yes	en	5	5	5	15	0	html tags in description
7	Romance studies	Advanced training course	yes	it	5	5	5	15	0	short description
8	Institute for Catholic Theology	Medieval Encounters between Christianity and Islam	no	en	5	5	5	14	1	no description
9	Institute for Computer Science	Lecture tutorial: System Ecozones and Living Beings	yes	de	5	5	5	15	0	long description; description contains literature references
10	Institute of Art History	Travel pictures. Artist journeys since the early modern period.	yes	de	5	5	5	14	1	
11	German studies	Information Structure, Syntax and Text Coherence (SW3, WP BA)	yes	de	5	5	5	14	1	
12	Institute for History	The French Revolution	yes	de	5	5	5	11	4	
13	Institute of Art History	Form as standard? Social utopias and individual living concepts.	yes	de	5	5	5	12	3	long description
14	Institute for Computer Science	Graph algorithms	yes	de	5	5	5	14	1	
15	English Studies/Art/Art Education	Making Reality: Art, Storytelling, Technology, and the Environment	yes	en	5	5	5	15	0	description contains comment
16	Institute of Psychology	Human-Computer Interaction	yes	de	5	5	5	14	1	
17	Economic Sciences	WiWi-B-01015-MA: Fundamentals of Marketing	yes	de	5	5	5	12	3	short description
18	Social Sciences	Introduction to spatial data visualization and analysis in political science (BA und MA)	yes	en	5	5	5	14	1	
19	Institute for Environmental Systems Research	Interdisciplinary seminar on applied research methods	yes	en	5	5	5	15	0	
20	Romance studies	Style and Expressive Modalities	yes	es	5	5	5	15	0	
21	Romance studies	Grammar 1	yes	fr	5	5	5	15	0	

Figure 26: Overview test set data.

Table 24: Result structure regarding model types and overlap.

Result structure (model_types)	n	%
results per model_type "tfidf"	110	
results per model_type "wikifier"	110	
results per model_type "doc2vec"	110	
results sum	330	100.0
results merged sum	305	92.4
results double sum	25	7.6

and stored. In total, 330 results (100%) were delivered by the search engine of which 25 were double (7.6%) and resulted in a usable result count of 305 (92.4%).

Table 25 shows the language mix of the test set courses. In total, there are five different languages included. 54.5% of the courses are in German, 31.8% in English and 4.5% each in Italian, French and Spanish. This roughly represents the language structure of the courses offered at UOS.

Table 25: Quantities and percentages of course languages.

Languages	n	%
German	12	54.5
English	7	31.8
Italian	1	4.5
French	1	4.5
Spanish	1	4.5

Table 26 shows the departments of the University of Osnabrück to which selected courses of the test set belong. In total, the test set covers courses from 15 departments. Most courses come from the Institute of Educational Sciences with 26.7% and the second most from the Romance Studies department with 20.0%. 13.3% of the courses are extracted from the Institute for Computer Science and 13.3% from the Institute of Art History. One course (6.7%) from each of 11 other subject areas is included in the test set. The distribution of the departments is partly representative for the UOS, e.g. the educational science share.

Table 26: Quantities and percentages of the test set faculties structure.

Department of the University of Osnabrück	n	%
Institute for Educational Science	4	26.7
Romance studies	3	20.0
Institute for Computer Science	2	13.3
Institute of Art History	2	13.3
Anglistics	1	6.7
Economic Sciences	1	6.7
English Studies/Art/Art Education	1	6.7
German studies	1	6.7
Institute for Catholic Theology	1	6.7
Institute for Environmental Systems Research	1	6.7
Institute for History	1	6.7
Institute of Psychology	1	6.7
LE Cognitive Science	1	6.7
Neurobiology	1	6.7
Social Sciences	1	6.7

A.3 Recommendation Engine state and language structure

Table 27 shows the quantities and percentages of indexed OER per languages for Discovery Pilot 1, Discovery Pilot 2 and the differences. Most of the indexed resources are in English with an quantity of $n = 85057$ (75% of the total amount) and increased by 17% since the last pilot.

Figure 27 shows the differences of the indexed contents for the top 7 languages. The largest increases were in English with 12,447 items (+17%) and Slovenian with 10,004 items (+312%). Further growth was achieved in German content with +1407 items (188%). For the first time, 280 Polish contents were added, which represents an increase of 100%. Italian content was reduced by 100 (-1%) and Spanish content by 331 (-8%).



Table 27: Structure of indexed OER per language for pilot 1, pilot 2 and their differences in numbers and percentages (X5gon Database).

Lang Code	Discovery n	Pilot 2 %	Discovery n	Pilot 1 %	diff	
					n	%
en	85057	75.961	72610	82.236	12447	17
sl	13119	11.716	3115	3.528	10004	321
it	6908	6.169	7008	7.937	-100	-1
es	3765	3.362	4096	4.639	-331	-8
de	2155	1.925	748	0.847	1407	188
pl	280	0.250			280	100
fr	192	0.171	196	0.222	-4	-2
el	179	0.160	179	0.203	0	0
ca	117	0.104	142	0.161	-25	-18
zh	85	0.076	85	0.096	0	0
da	29	0.026	29	0.033	0	0
ja	21	0.019	21	0.024	0	0
pt	20	0.018	20	0.023	0	0
la	15	0.013	15	0.017	0	0
hr	5	0.004	5	0.006	0	0
ru	5	0.004	5	0.006	0	0
bs	4	0.004	4	0.005	0	0
ml	3	0.003	3	0.003	0	0
sr	3	0.003	3	0.003	0	0
sa	2	0.002	2	0.002	0	0
eu	2	0.002	2	0.002	0	0
id	2	0.002			2	100
hu	2	0.002			2	100
nl	2	0.002	2	0.002	0	0
gl	1	0.001	1	0.001	0	0
kk	1	0.001	1	0.001	0	0
ia	1	0.001	1	0.001	0	0
id			2	0.002	-2	-100

Figure 28 shows the language composition of the indexed contents as a pie chart (top 7 languages). It can be clearly seen that the largest part of the content is in English and corresponds to about 3/4 (76.0%). Slovenian content has a relative share of 11.7% and Italian 6.2%. Besides Spanish content, which corresponds to 3.4%, and German content (1.9%), all other languages are rather under-represented (less than or equal to 0.25%).

Difference: content quantity for top 7 languages
(Discovery Pilot 1 to Discovery Pilot 2)

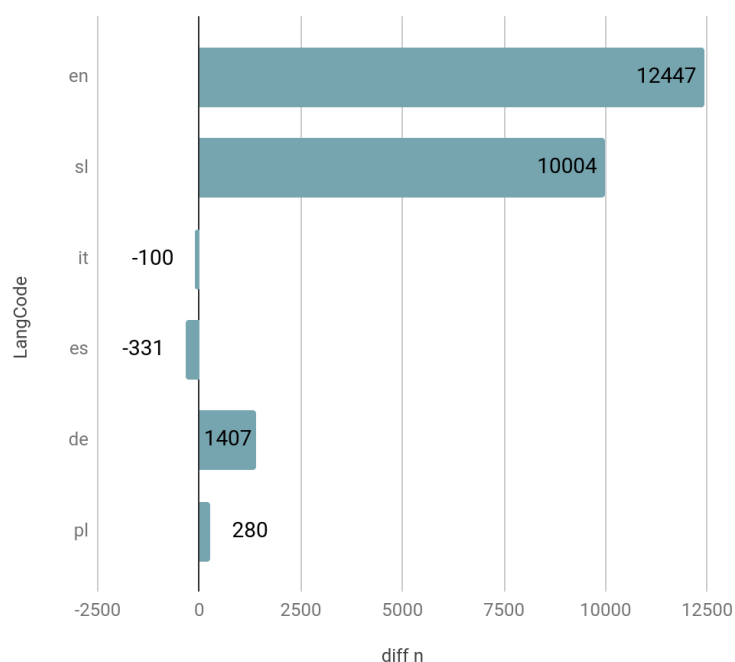


Figure 27: Bar chart of content quantities for indexed language (top 7).

Content Language Structure (2020-01-22)

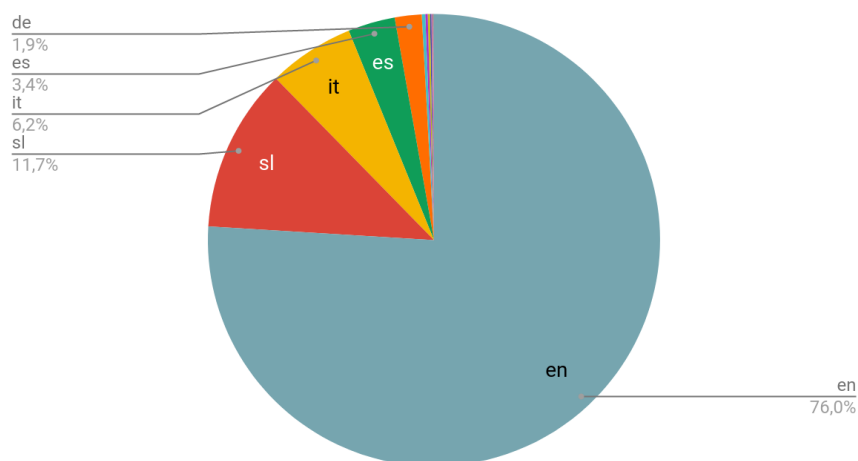


Figure 28: Diagram of the language structure of indexed content/OER.

B MediaUPV with multilingual subtitles as of June 2020

In a broad sense, the MediaUPV repository is a professional UPV service for the creation, storage, management and open dissemination of educational videos [54, 55]. Launched in 2007, it was initially designed for UPV lecturers to produce high-quality short video recordings at dedicated UPV studios, with the aim of supporting blended learning through prerecorded “knowledge pills”. These recordings, usually referred to as *poliMedias*, have also served as the main back-end video service for the UPV to provide MOOCs [56], especially as an edX member since 2014 [57]. In this respect, it is worth noting that UPV has become one of the most renowned MOOC providers in Spanish, with more than 85 MOOCs and 290 editions already completed, more than 2.3 million enrollments, and two of the 100 most popular online courses of all time [58]. Apart from poliMedias, MediaUPV has been expanded to include homemade videos produced by students and lecturers themselves, known as *poliTubes*, which are uploaded to it in much the same way as in YouTube. Finally, since joining the Opencast consortium in 2011, UPV has deployed lecture capture technology to 84 locations from which more than 600 hours per year are being recorded and added to MediaUPV for their distribution to students only through a Sakai LMS [59, 60].

Although MediaUPV comprises diverse kinds of educational videos, this study focuses only on poliMedias due to their relevance to X5gon and simplicity in terms of duration, speakers and audio quality. As indicated above, they are produced at dedicated UPV studios which, in brief, are just low-cost video production (4x4 metre) rooms equipped with a white backdrop, video camera, capture station, pocket microphone, lighting and AV equipment including a video mixer and an audio noise gate (Figure 29, left). After choosing day and time of an appointment by an online booking system, the lecturer comes to a poliMedia studio with slides and delivers her/his presentation in front of the video camera, which is captured and synchronously embedded in real-time at the bottom-right corner of the computer’s video output. Then, after metadata annotation, review and approval by the lecturer, the resulting poliMedia is uploaded to MediaUPV (see example in Figure 29, right).

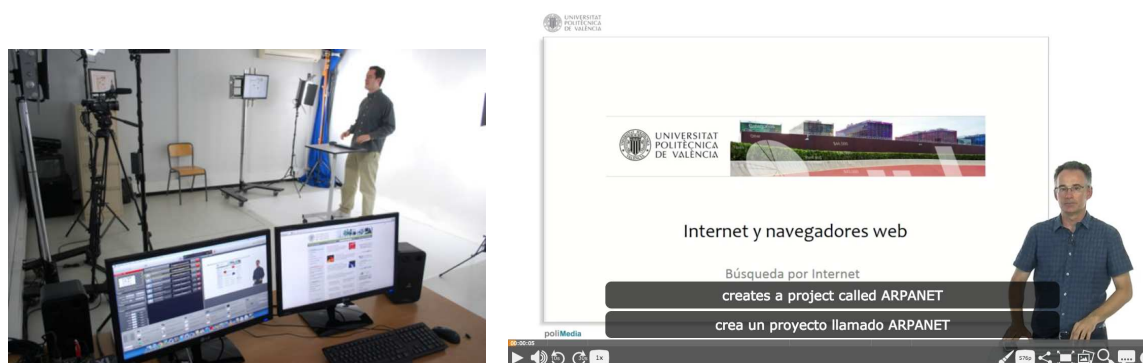


Figure 29: A poliMedia studio (left) and example (right).

Supported by the UPV’s *Docència en Xarxa (DeX)* stimulus plan for online teaching, the number of poliMedias uploaded to MediaUPV has been steadily increasing since 2007, up to 44096 videos and a total of 10601 recording hours in June 2020. As with face-to-face teaching sessions, the vast majority of poliMedias are produced in Spanish though, as shown in Table 28, they are also produced, to a much lesser extent, in Catalan (also known as Valencian in the Valencian Community) and English. In this regard, the UPV has recently approved an ambitious plan to promote multilingual teaching for the period 2020–2023 in which Catalan and English are specifically identified as top priorities for support [61, pp 120–144]. On the one hand, Catalan is an official yet minority language in the Valencian Community, and thus its protection is seen not only as an appreciation of cultural diversity,

but also an obligation to reduce discrimination on the grounds of language at the UPV. The case of English, on the other hand, is totally different. Increasing its use as a teaching language is clearly needed to strengthen UPV's internationalization and competitiveness. It goes without saying that, for this plan to succeed, it will be good to have accurate and cost-effective means to fully convert basic (monolingual) poliMedias into trilingual learning objects.

Table 28: Number of poliMedia videos and hours in Spanish, Catalan and English.

Language	Videos		Hours	
	No.	%	No.	%
Spanish	38172	87	9451	89
Catalan	1333	3	232	2
English	4591	10	918	7
Total	44096	100	10601	100

Table 29 shows the number of lecturers producing poliMedias in each of the seven possible combinations of Spanish (es), Catalan (ca) and English (en): three of them monolingual (es, ca, en), three bilingual (es-ca, es-en, ca-en) and the trilingual case es-ca-en. Note that the percentages of monolingual, bilingual and trilingual lecturers are 91.9, 7.6 and 0.5, respectively. This means that a great majority of lecturers are producing poliMedias in a single language, Spanish in most cases, to support their face-to-face teaching sessions. Also worth noting is the fact that the number of lecturers producing poliMedias in English (872) is roughly 4 times that of poliMedias in Catalan (212), yet both languages account for a similar percentage of the total academic offer [61, pp 120–144]. This is because all Catalan-speaking learners are highly proficient in Spanish, and thus poliMedias in Spanish are also often used to support blended learning for Catalan-language groups. Needless to say, promoting multilingualism (in the UPV) means that all supported languages must be treated equally with regard to available resources.

Table 29: poliMedia lecturers for Spanish (es), Catalan (ca), English (en), bilingual combinations (es-ca, es-en, ca-en) and the trilingual case es-ca-en.

	Monolingual			Bilingual			Trilingual	Total
	es	ca	en	es-ca	es-en	ca-en	es-ca-en	
No.	2126	152	656	43	199	2	15	3193
%	66.6	4.8	20.5	1.3	6.2	0.1	0.5	100.0
Total (%)	91.9			7.6			0.5	100.0

The MediaUPV repository is a good example of how OER repositories are evolving in terms of size and complexity, especially at the linguistic level. This is why, (the poliMedia part of) it was chosen as a case study in X5gon. Indeed, before the beginning of X5gon, poliMedia-adapted ASR/MT systems were already integrated into the MediaUPV production workflow to enrich all poliMedias with raw multilingual subtitles [42]. Prior to X5gon, however, it was felt that post-editing raw subtitles was still needed in many cases, and thus a user-friendly tool for reviewing was also integrated into the production workflow [62, 63, 64, 65]. Being part of this workflow, subtitle post-editing was supported by the DeX stimulus plan, allowing each poliMedia to be reviewed not only by its author, but also by non-authors (e.g. users), with the author's approval prior to publication. Although this post-editing approach worked (and still works) well, poliMedias have been more and more published with no subtitle post-editing at all due to the increasing accuracy of new ASR/MT systems. In fact, our latest results show that we have now reached the point at which raw subtitles are often good enough for direct

publication. To be more precise, Table 30 provides some figures on the quality of our current ASR and MT systems: X5gon’s M40 systems for Spanish and English (Section 3) and internal systems for Catalan. For comparison, Table 30 also provides analogous figures for general-purpose systems now commercially available from Google (*Google Cloud Speech-To-Text*, *standard* and *enhanced* if available; and *Google Translate*), and their relative value with respect to those built at the UPV ($\Delta\%$).

Table 30: WER/BLEU scores provided by UPV and Google ASR/MT systems on poliMedias (es=Spanish, ca=Catalan, en=English, “es \Rightarrow ca”=“Spanish to Catalan”, etc.)

Systems	ASR WER (%)				MT BLEU (%)			
		es	ca	en		es \Rightarrow ca	es \Rightarrow en	ca \Rightarrow es
UPV	ASR	8.3	12.6	12.0	MT	84.4	34.1	90.3
Google	S2T standard	19.9	31.9	36.1	Translate	81.5	36.8	87.7
	S2T enhanced	n/a	n/a	13.3				
	$\Delta\%$	139.8	153.2	10.8		$\Delta\%$	-3.4	7.3

For the analysis of results in Table 30, it should be pointed out first that there are no simple, error-free rules to decide, from WER and BLEU scores, whether raw subtitles are publishable or not. On the contrary, being a derivative work of an educational video owned by a lecturer, (raw) subtitles can be approved for publication only with the owner’s consent and after review if desired. This is indeed the way in which publication consent has been sought for poliMedias and, in doing so, it was soon realized that little or none subtitle post-editing was actually done as ASR/MT accuracy improved. To be precise, this was clearly observed for source subtitles produced by ASR systems with WER figures below 20%, as well as for target subtitles generated by MT systems with BLEU scores above 35% [41, 5]. With these thresholds in mind, we see that UPV’s raw subtitles are good enough for direct publication in all cases, except perhaps in the case of Spanish to English translation, whose BLEU score is slightly below 35%. In fact, the WER and BLEU scores for Spanish and Catalan, around 10% WER and above 84% BLEU, are far better than these thresholds. In comparison, if we look at Google’s results and their relative value, we see that Google’s general-purpose systems are also fairly good, especially for MT and English ASR, though they are clearly behind our task-adapted ASR systems.

References

- [1] Erik Novak. X5gon deliverable 2.2: Final Server Side Platform. Technical report, JSI, August 2019. <https://www.x5gon.org/science/deliverables>.
- [2] Colin de la Higuera and Walid Ben Romdhane. X5gon deliverable 3.2: Learning Analytic Engine 2.0. Technical report, NA, August 2019. <https://www.x5gon.org/science/deliverables>.
- [3] Colin de la Higuera, Walid Ben Romdhane, and Victor Connes. X5gon deliverable 3.3: Learning Analytic Engine 3.0. Technical report, NA, February 2020. <https://www.x5gon.org/science/deliverables>.
- [4] Alex Pérez, Javier Iranzo, and Alfons Juan. X5gon deliverable 5.2: Second report on piloting. Technical report, UPV, August 2019. www.x5gon.org/science/deliverables.
- [5] Javier Jorge and Alfons Juan. X5gon deliverable 3.5: Final support for Cross-lingual OER. Technical report, UPV, February 2020. www.x5gon.org/science/deliverables.
- [6] Álex Pérez, Joan Albert Silvestre, and Alfons Juan. X5gon deliverable 5.1: First report on piloting. Technical report, UPV, August 2018. www.x5gon.org/science/deliverables.
- [7] Dhanashri Ingale and Manali M. Kshirsagar. A review on an approach to infer user search goals for optimize result. In *2016 World Conf. on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave)*, pages 1–4, 2016.
- [8] J. Beus. Klickwahrscheinlichkeiten in den Google SERPs. Retrieved 26 April 2019, from SISTRIX website. <https://www.sistrix.de/news/klickwahrscheinlichkeiten-in-den-google-serps/>, 2015.
- [9] Elaine Tan and Nick Pearce. Open education videos in the classroom: exploring the opportunities and barriers to the use of youtube in teaching introductory sociology. *Research in Learning Technology*, 19, 2011.
- [10] H. Jung, H. V. Shin, and J. Kim. Dynamicslide: Reference-based interaction techniques for slide-based lecture videos. In *The 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings*, pages 23–25, 2018.
- [11] K. Yadav, A. Gandhi, A. Biswas, K. Shrivastava, S. Srivastava, and O. Deshmukh. Vizig: Anchor points based non-linear navigation and summarization in educational videos. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 407–418, 2016.
- [12] Carla F Griggio, Nam Giang, Germán Leiva, and Wendy E Mackay. The uist video browser: Creating shareable playlists of video previews. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 59–60, 2016.
- [13] Jeremy C Green, Taha Aziz, Juliane Joseph, Angad Ravanam, Sobia Shahab, and Luke Straus. Youtube enhanced case teaching in health management and policy. *Health professions education*, 4(1):48–58, 2018.
- [14] Chareen Snelson. Mapping youtube” video playlist lessons” to the learning domains: Planning for cognitive, affective, and psychomotor learning. In *Society for Information Technology & Teacher Education International Conference*, pages 1193–1198. Association for the Advancement of Computing in Education (AACE), 2010.



- [15] Anja Nylund Hagen. The playlist experience: Personal playlists in music streaming services. *Popular Music and Society*, 38(5):625–645, 2015.
- [16] Sharon McDonald, Helen M Edwards, and Tingting Zhao. Exploring think-alouds in usability testing: An international survey. *IEEE Transactions on Professional Communication*, 55(1):2–19, 2012.
- [17] Crystal A Kalinec-Craig, Jaime M Diamond, and Jeffrey Shih. A playlist as a metaphor for engaging in a collaborative self-study of mathematics teacher educator practices. *Studying Teacher Education*, 16(3):345–363, 2020.
- [18] Javier Iranzo, Álex Pérez, Jorge Civera, Albert Sanchis, and Alfons Juan. X5gon deliverable 3.4: Early support for Cross-lingual OER. Technical report, UPV, August 2019. <https://www.x5gon.org/science/deliverables>.
- [19] Pau Baquero-Arnal, Javier Jorge, Adrià Giménez, Joan Albert Silvestre-Cerdà, Javier Iranzo-Sánchez, Albert Sanchis, Jorge Civera, and Alfons Juan. Improved Hybrid Streaming ASR with Transformer Language Models. In *Proc. of Interspeech 2020*, pages 2127–2131, October 2020. <http://dx.doi.org/10.21437/Interspeech.2020-2770>.
- [20] Javier Jorge, Adrià Giménez, Joan Albert Silvestre-Cerdà, Jorge Civera, Albert Sanchis, and Alfons Juan. Live Streaming Speech Recognition using Deep Bidirectional LSTM Acoustic Models and Interpolated Language Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021. (submitted).
- [21] Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. Re-translation versus Streaming for Simultaneous Translation. *arXiv preprint arXiv:2004.03643*, 2020.
- [22] Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proc. of ACL*, pages 3025–3036, 2019.
- [23] Colin Raffel, Minh-Thang Luong, Peter J. Liu, Ron J. Weiss, and Douglas Eck. Online and linear-time attention by enforcing monotonic alignments. In *Proc. of ICML*, volume 70, pages 2837–2846. PMLR, 2017.
- [24] Xutai Ma, Juan Miguel Pino, James Cross, Liezl Puzon, and Jiatao Gu. Monotonic Multihead Attention. In *Proc. ICLR 2020*. OpenReview.net, 2020.
- [25] Maha Elbayad, Laurent Besacier, and Jakob Verbeek. Efficient Wait-k Models for Simultaneous Machine Translation. In *Proc. Interspeech 2020*, pages 1461–1465, 2020.
- [26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Proc. of NIPS*, pages 5998–6008, 2017.
- [27] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.



- [28] Roei Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [29] Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019, 2019.
- [30] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy, July 2019. Association for Computational Linguistics.
- [31] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proc. of NAACL-HLT: Demonstrations*. ACL, 2019.
- [32] Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).
- [33] Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. Report on the 11th IWSLT evaluation campaign, IWSLT 2014. In *Proc. of IWSLT*, pages 2–17, 2014.
- [34] Javier Jorge, Adrià Giménez, et al. LSTM-Based One-Pass Decoder for Low-Latency Streaming. In *Proc. of 2020 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7814–7818, 2020.
- [35] W3C. Web Content Accessibility Guidelines (WCAG) 2.1. www.w3.org/TR/WCAG21, 2018.
- [36] Robert Godwin-Jones. In a World of SMART Technology, Why Learn Another Language? *Journal of Educational Technology & Society*, 22(2):4–13, 2019.
- [37] Carolien A.N. Knoop van Campen, Eliane Segers, and Ludo Verhoeven. Effects of audio support on multimedia learning processes and outcomes in students with dyslexia. *Computers & Education*, 150, 2020.
- [38] Erin K. Chiou, Noah L. Schroeder, and Scotty D. Craig. How we trust, perceive, and learn from virtual humans: The influence of voice quality. *Computers & Education*, 146, 2020.
- [39] Jonathan Shen, Ruoming Pang, et al. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In *Proc. of 2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783, 2018.
- [40] Yu Zhang, Ron J. Weiss, et al. Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning. In *Proc. of Interspeech 2019*, pages 2080–2084, 2019.
- [41] Juan D. Valor-Miró, Pau Baquero-Arnal, Jorge Civera, Carlos Turró, and Alfons Juan. Multilingual videos for MOOCs and OER. *Journal of Educational Technology & Society*, 21(2):1–12, 2018.



- [42] S. Piqueras, M. A. Del-Agua, A. Giménez, J. Civera, and A. Juan. Statistical Text-to-Speech Synthesis of Spanish Subtitles. In *Proc. of the 2nd Int. Conf. on Advances in Speech and Language Technologies for Iberian Languages (IberSpeech)*, volume 8854, pages 40–48. 2014.
- [43] Yi Ren, Yangjun Ruan, et al. FastSpeech: Fast, Robust and ControllableText to Speech. In *Proc. of the 33rd Conf. on Neural Information Processing Systems (NeurIPS)*, 2019.
- [44] Wei Ping, Kainan Peng, et al. Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. In *Proc. of the Sixth Int. Conf. on Learning Representations (ICLR)*, 2018.
- [45] Aäron van den Oord, Sander Dieleman, et al. WaveNet: A Generative Model for Raw Audio. arXiv preprint arXiv:1609.03499.pdf, 2016.
- [46] Nal Kalchbrenner, Erich Elsen, et al. Efficient Neural Audio Synthesis. In *Proc. of the 35th International Conference on Machine Learning (ICML 2018)*, volume PMLR 80, pages 2410–2419, 2018.
- [47] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). Centre for Speech Technology Research (CSTR), University of Edinburgh, 2019.
- [48] Rayhane Mama. Tacotron-2. github.com/Rayhane-mamah/Tacotron-2, 2018.
- [49] Ollie McCarthy. Wavernn. github.com/fatchord/WaveRNN, 2018.
- [50] Melvyn J. Hunt. Figures of merit for assessing connected-word recognisers. *Speech Communication*, 9(4):329–336, 1990.
- [51] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002.
- [52] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proc. of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231, 2006.
- [53] ITU-T. *ITU-T Recommendation P.85: A method for subjective performance assessment of the quality of speech voice output devices*. Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T), 1994. <https://www.itu.int/rec/T-REC-P.85-199406-I/en>.
- [54] Carlos Turró, Miguel Ferrando-Bataller, Jaime Busquets, and Aristóteles Cañero. Polimedia: a system for successful video e-learning. In *Proc. of the EUNIS Annual Congress*, 2009.
- [55] MediaUPV. The MediaUPV repository. <https://media.upv.es>, 2020. Retrieved on June 2020.
- [56] UPVX. UPVX: The MOOC initiative at the UPV. <https://www.upvx.es>, 2020. Retrieved on June 2020.
- [57] UPValenciaX. UPValenciaX: UPV as an edX member. <https://www.edx.org/school/upvalenciax>, 2020. Retrieved on June 2020.
- [58] ClassCentral. The 100 Most Popular Online Courses of All Time (2020). <https://www.classcentral.com/report/most-popular-online-courses>, 2020. Retrieved on June 2020.



- [59] Carlos Turró, Ignacio Despujol, Aristóteles Cañero, and Jaime Busquets. Deployment and Analysis of Lecture Recording in Engineering Education. In *Proc. of 2014 IEEE Frontiers in Education Conference (FIE)*, pages 1–5, 2014.
- [60] Opencast. Opencast. <https://opencast.org>, 2020. Retrieved on June 2020.
- [61] BOUPV20. Official Bulletin of the UPV. <http://hdl.handle.net/10251/145577>, 2020. Retrieved on June 2020 (in Catalan and Spanish).
- [62] J. A. Silvestre-Cerdà, A. Pérez, M. Jiménez, C. Turró, A. Juan, and J. Civera. A System Architecture to Support Cost-Effective Transcription and Translation of Large Video Lecture Repositories. In *Proc. of 2013 IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC)*, pages 3994–3999, 2013.
- [63] Alejandro Pérez, Joan Albert Silvestre-Cerdà, Juan Daniel Valor-Miró, Jorge Civera, and Alfons Juan. MLLP Transcription and Translation Platform. In *Proc. of the 10th European Conf. on Technology Enhanced Learning (EC-TEL)*, 2015.
- [64] Juan D. Valor-Miró, Joan A. Silvestre-Cerdà, Jorge Civera, Carlos Turró, and Alfons Juan. Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories. *Speech Communication*, 74:65–75, 2015.
- [65] Juan D. Valor-Miró, Joan A. Silvestre-Cerdà, Jorge Civera, Carlos Turró, and Alfons Juan. Efficient Generation of High-Quality Multilingual Subtitles for Video Lecture Repositories. In *Proc. of the 10th European Conf. on Technology Enhanced Learning (EC-TEL)*, pages 485–490, 2015.

