UTTER

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – Report on first set of FSTP projects

Nature	Report	Work Package	WP1
Due Date	31/10/2024	Submission Date	31/10/2024
Main authors	Wilker Aziz (UVA)		
Co-authors			
Reviewers	Maryam Hashemi		
Keywords	survey, languages, resources		
Version Control			
v1.0	Status	Final	31/10/2024

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

List of Figures

Abstract

Within UTTER we have allocated EUR 895 206.00 to run an FSTP programme, which is a key component of UTTER's impact. We have split our programme in two separate calls for project proposals, one running in 2023/2024, the other running in 2024/2025. In this document (D1.2), we describe our first call for FSTP project proposals and its outcomes. Our first call attracted 54 submissions, of which 8 were selected for funding. All 8 projects started in January 2024 and concluded successfully in September 2024. For completeness, Appendix B contains the final reports from the 8 project teams. At the time of writing, a second and final call is ongoing, that call will be covered in D1.3 (M36).

1 First Call

1.1 Infrastructure

To ensure full compliance with our data management plan (D1.1),[1] we managed our FSTP call (both submission and review process) via a UVA-hosted instance of HotCRP[2]—an open-source software for managing conference review processes. Our site[3] received submissions from July 31, 2023 to October 15, 2023.[4]

We hosted the call documentation on our website and further linked it from the European commission Funding and tender opportunities website.

For communication, we created a UVA-hosted mailing list (utter-fstp@list.uva.nl).

1.2 Dissemination

In accordance with the GA (MS6), we posted the FSTP call on our own website[5] on July 6, 2023 and on the European commission Funding and tender opportunities website[6] on July 27, 2023. We further advertised the FSTP call in a number of ways:

- On July5 , 2023 UTTER hosted its 1st User Day,[7] where we advertised our (then upcoming) 1st FSTP call, explaining its objectives and sharing key facts;[8]
- On September 15, 2023 UTTER joined a remote dissemination event targetting contemporary FSTP calls organised by Sploro;[9]
- UTTER's social media channels (e.g., Twitter and Linkedin);[10]
- the PIs' own networks.[11]

1.3 Call documentation

The complete call is described in a core Call Documentation[12] document along with 4 annexes:

- **A1 Guide for applicants.** A document with detailed instructions and screenshots to guide applicants through the submission process;[13]

[1] https://projectnetboard.absiskey.com/viewdocument/5078a3-3d1fd5-a89ee2-cbe72c-000033
[2] https://hotcrp.com
[3] https://utter-fstp.science.uva.nl
[4] At the time of writing (October 2024), the site is still live as we are currently running our 2nd FSTP call. We have stored in our UVA-hosted GitLab instance, a complete snapshot of the website's database prior to launching the 2nd call.
[5] https://he-utter.eu/#fstp
[6] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/competitive-calls-cs/3722
[7] A recording of which is accessible from this URL https://www.youtube.com/watch?v=3Bm_3C9HrP0
[8] https://projectnetboard.absiskey.com/viewdocument/367924-3d674c-5727c8-c8dfc1-000021
[9] https://sploro.eu/cascade-funding-opportunities-october-2023/
[10] e.g., https://x.com/UTTERProject/status/1691031765530386432
[11] e.g.,https://x.com/bazril/status/1688468973522755584
[12] https://projectnetboard.absiskey.com/viewdocument/3fa72e-46ba65-65c53f-7f3a64-000029
[13] https://projectnetboard.absiskey.com/viewdocument/8c84ee-dc0d51-70cadb-c05289-000028

2023-07-06

HE-UTTER FINANCIAL SUPPORT FOR THIRD PARTIES (FSTP) CALL

Funding opportunity for research organization and SMEs – development and application of deep models for extended reality

Our Horizon Europe project, **UTTER**, invites project proposals from research organizations and SMEs to develop and/or pilot applications of large pretrained language models with a focus on enabling human-human and human-machine interaction. Successful applications will receive **up to € 60 000** each, and run for **6-9 months**. The call closes on **October 15, 2023**.

Below is an exhaustive compilation of essential documentation pertaining to the Submission & Evaluation process:

- FSTP- Key Facts
- UTTER- Call documentation
- UTTER- Grant Agreement
- UTTER- Consortium Agreement
- A1- guide for applicants
- A2- third party agreemetns
- A3- project proposal template
- A4- Evaluation criteria

Proposals are to be submitted via UTTER's proposal management portal at: https://utter-fstp.science.uva.nl

Please check out these two video recordings showcasing our project prototypes:

- Demo-José Sousa
- Demo-Laurent Besacier

Figure 1: Screenshot of the call documentation package, as disseminated through our website on July 6, 2023.

- **A2 Third party agreement.** The agreement that awardees would have to sign in order to execute their project under our funding scheme;[14]
- **A3 Project proposal template.** A docx proposal template, with the required fields and guidelines of how to fill them in;[15]
- **A4 Evaluation criteria.** The description of the evaluation process and criteria.[16]

In addition, we shared UTTER's GA and CA. For ease of reference, we also prepared a key facts sheet, mostly used for dissemination.[17] See Figure 1 for a screenshot of the relevant entry on our website.

1.4 Key Parameters

Objectives. Develop and/or pilot applications using XR models (i.e., pre-trained neural network models adaptable to a large variety of forms of expression, interaction, languages, domains, styles and intent) in new sectors, with a focus on enabling new types of human-human and human-machine interaction. Examples of welcome project objectives include:

- Improving or demonstrating efficiency of XR model inference;
- Improving or demonstrating efficiency of XR model training;

[14] https://projectnetboard.absiskey.com/viewdocument/83d5ba-eba9a2-50d716-f433cd-000024
[15] https://projectnetboard.absiskey.com/viewdocument/307c36-90b9a6-745187-ebcd62-000026
[16] https://projectnetboard.absiskey.com/viewdocument/d2c134-3a90c9-0e67b8-0f4e7d-000027
[17] https://projectnetboard.absiskey.com/viewdocument/367924-3d674c-5727c8-c8dfc1-000021

- Designing interfaces for usability;
- Extending XR models to new languages, domains or modalities;
- Applying XR models to new tasks;
- Building resources for XR models;
- Evaluation of XR models.

Proposals.

- Maximum budget per project: 60,000 euro
- Project duration: 6–9 months
- Applicant: SME or research organisation from a Horizon Europe eligible country

Project execution.

- Development
- Dissemination

1.5 Proposal Template

The templated collected the following information, in structured format.

- Project identification
- Applicant identification
- Project description
- Project team
- Budget
- Ethics self-assessment
- Detailed Budget
- Consent to process personal data
- Declaration of Honour

1.6 Evaluation

1.6.1 Programme committee

We organised a programme committee (PC) to carefully assess the proposals. Structure of PC:

- Coordinator: Wilker Aziz (UTTER/UVA).
- Manager: Maryam Hashemi (UTTER/UVA). Within our process the manager assists with formal checks as well as monitoring the mailing list, arranging payment for external reviewers, and various other tasks.
- Chairs: UTTER research personnel. Within our process a PC chair will a) assess proposals for adequacy, b) perform full qualitative reviews, c) invite external reviewers to contribute full qualitative reviews, and d) contribute to final decisions.
- External reviewers: experts with a PhD (or senior PhD candidates) who are not part of UTTER nor in active collaboration with UTTER personnel.

We refer to Coordinator and Chairs collectively as the Pilot Board.

1.6.2 Conflict of Interest (CoI)

There are two types of CoI that are relevant for our process:

- The applicant *is* in active collaboration with UTTER partners. We only welcome proposals that are free of this kind of CoI. Applicants self-declare this form of CoI, as part of the submission form, and the Coordinator verifies that no such CoI went unnoticed as part of a formal requirements check.
- The applicant *has been* in some form of collaboration with UTTER partners.
  We treat this as the regular kind of CoI in conference reviewing processes, namely, this form of CoI rules out the UTTER partner in conflict as a Chair for the proposal.

1.6.3 Criteria

We have three sets of evaluation criteria, all of which are clearly described in the call documentation Annex 4. They are summarised next.

Formal requirements. Every proposal must comply with a number formal requirements. These mostly ensure:

- the proposal is written in English, submitted on time using the proposal template;
- the applicant signed a DoH;
- the applicant's legal status fits the call;
- the applicant is legally established in a Horizon Europe eligible country;

- there's no CoI (i.e., no active collaboration with UTTER partners);
- the applicant filled in an ethical self-assessment.

Formal requirements are treated as Yes/No criteria, checked by the Manager and the Coordinator. All formal requirements must be met.

Eligibility criteria. Proposals that meet formal requirements are then assessed for their eligibility to the call, which we prescribed in terms of the following:

- *Relevance.* Does the proposal address the objectives of the call?
- *Uniqueness.* Does the proposal break new ground?
- *Completeness.* Does the proposal include both project phases (i.e., Development and Dissemination)?

These are treated as Yes/No questions, an eligible proposal meets all 3 criteria. Our goal is to filter out clear cases where the applicant misunderstands the goals of the call, or the proposal is obviously redundant (i.e., with clear known examples to support this assessment), or the proposal does not address (or address arguably superficially) the two project phases. Eligibility criteria are assessed by Chairs aiming at having high recall (i.e., we prefer to mark a borderline proposal as eligible than to mark a potentially eligible proposal as ineligible).

Qualitative criteria. Finally, eligible proposals undergo a complete qualitative assessment. The critearia are listed below.

- *Objective fit (2).* Are the project goals and planned achievements in line with the overall objectives of UTTER? Is it likely that the project will deliver added value to UTTER?
- *Technical approach (2).* Are the planned activities feasible and facilitate the achievement of project outputs? Does the proposal push the boundaries of existing XR technology?
- *Business, Integration and Dissemination (BID) plan (3).* Is the business plan reasonable and ambitious? How well is the integration of project outputs planned? Are the dissemination and promotion activities planned adequately?
- *Budget adequacy (1).* Does the budget correspond to all planned activities and outputs?
- *Team (1).* Is the applicant's team capable of executing the project and delivering its outputs (in required time, quality and with estimated budget)?
- *Ethics (1).* Is the ethical self-assessment thoughtful and thorough? Does it provide convincing justification that the applicant will ensure the work will be done ethically?

In bracket, for each criteria, we indicate their weight towards a final numerical score used for ranking. To map the qualitative criteria to a numerical score, we use the scale listed in Table 1.

Score	Rubric
0	Not at all: The proposal fails to address the criterion or cannot be assessed due to missing or incomplete information
3	Limited: The criterion is inadequately addressed or there are significant weaknesses.
7	Good: The proposal addresses the criterion well, but some shortcomings are present.
10	Excellent: The proposal successfully addresses all relevant aspects of the criterion, any shortcomings are minor.

Table 1: Rating scale

1.6.4 Procedure

The evaluation of FSTP proposals was divided in phases, which we describe next. Figure 2 gives an overview of the process.

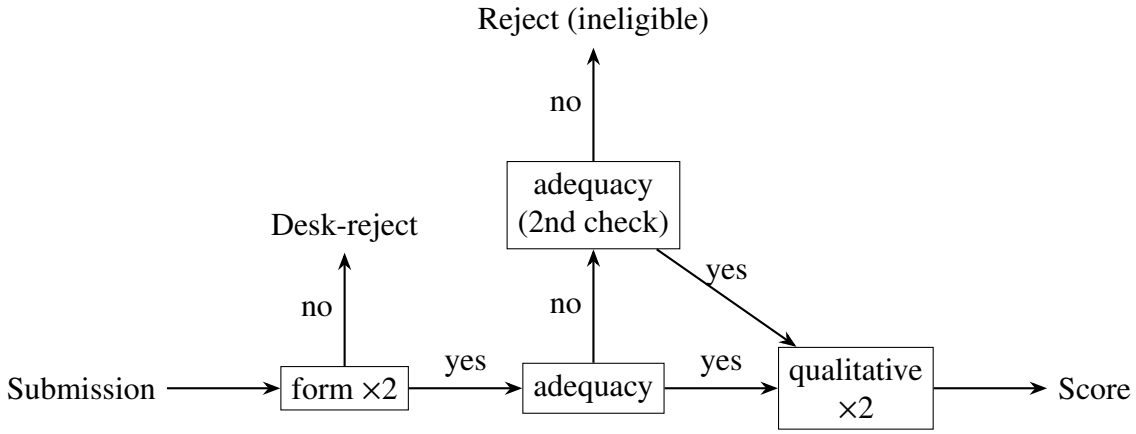


Figure 2: Overview of assessment phases (rectangles).

Formal requirements. In this phase, we assessed proposals on basic formal requirements. This check was conducted by both the Manager and the Coordinator.

Projects that did not comply with one or more formal requirements were marked for desk-rejection, with the following exceptions:

- Invalid or missing PIC
- Missing abstract
- Error in budget (e.g., wrong categories, typos, unexpected totals)

We assumed errors of these sort signalled a lesser degree of experience with FSTP calls and, given that these errors are easy to amend should the proposal be (conditionally) selected for funding, we opted for not desk-rejecting such cases.

Eligibility checks. Proposals that passed the formal check were then assessed for their eligibility to the call. The Coordinator assigned proposals to Chairs observing two criteria: a) topic alignment (where possible), b) lack of CoI. At this point the Coordinator's assignment was based on educated guesses, hence, this first assignment was adjusted based on the Chair's self-declared CoIs. After these adjustments, each Chair managed 4-5 proposals, all free of CoI. Proposals flagged as ineligible (*i.e.*, failing at one or more of the three criteria) were then independently assessed by a second Chair (again, free of CoI). The proposal is marked for rejection in case both Chairs agree on that outcome. In case of disagreement, we opted to regard the proposal as *eligible* and hence have it moved to the final phrase of assessment.

Qualitative assessment. Eligible proposals receive a full qualitative assessment by two independent evaluators (both free of CoI), a Chair and an External Reviewer (external reviewers self-declare CoI). Each evaluator evaluates all individual criteria using qualitative rubrics, which are then mapped to points. The points from all evaluators are averaged by criterion. Points by criterion are then multiplied by the criterion's weight and summed up in order to get the proposal's *overall score*. The Pilot Board can change the total number of points assigned to a proposal in the range of at most 30 points (up or down) of all the points the proposal received from the evaluators. The total overall score of an individual proposal is 130 points: maximum 100 points from evaluators + maximum 30 points from Pilot Board.

Decisions. We ranked the proposals for quality and selected those that received highest scores for funding subject to i) not selecting more than half of our total FSTP budget in our 1st call, and ii) not selecting proposals that were assigned widely different scores (e.g., should there be a clear top cluster followed by one or more second-tier clusters). With (i) in mind, we hoped to find 6–8 strong proposals forming a clear top cluster.

1.6.5 Review Forms

We designed review forms on HotCRP with all necessary information for reviewers, as well as rubrics to guide their assessment. There are 3 forms:

- Formal Requirements – Appendix A Figure 5
- Adequacy to Call (that's the form for eligibility criteria) – Appendix Figure 6
- Qualitative Assessment – Appendix A Figures 7 (Key Parameters), 8 (Objective fit), 9 (Technical approach), 10 (BID), 11 (Team and Budget), 12 (Ethics and Comment for PC).

1.7 Outcome

Here we summarise the outcome of our review process:

- Submissions: 54 (see Figure 3 for an overview of where they were submitted from).
- Desk-reject due to formal requirements: 5
- Ineligible after first check: 17

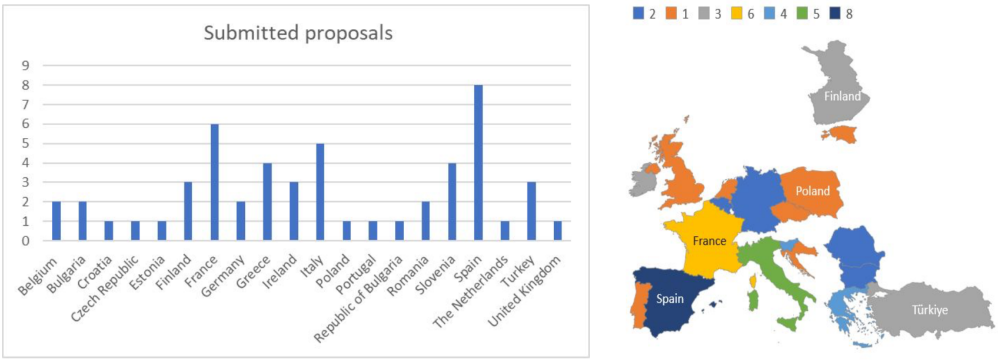


Figure 3: Proposals per country.

- Ineligible after second check: 16
- Qualitatively assessed: 33

Desk-rejected projects. 3 projects were desk-rejected because they failed to address the vast majority of required fields of the proposal template. 1 project was from an inelligible country, for this call. 1 proposal was a (likely accidental) duplicate—we only desk-rejected the copy (the original remained under consideration).

Ineligible projects. 15 projects were judged to fail along the *Relevance* dimension, 9 projects were judged to fail along the *Uniqueness* dimensions, and 1 project was judged to fail along the *Completeness* dimension. One project was flagged as potentially ineligible in the first check and then considered potentially eligible in a second check—this was the only case were 2 Chairs disagreed, this project was then treated as eligible. These are the most frequent criticism that Chairs identified to support their assessment: *wrong XR / confused XR for VR* (10), *unclear impact of XR* (8), *unclear goals* (7), *unrealistic dependencies* (4).

Selected for funding. We selected the top 8 proposals for funding, as that amounts to about half of the total budget allocated for FSTP calls in UTTER and those 8 proposals formed a cluster reasonably separated from the rest (see Figure 4).

Table 2 summarises the total funding requested.

	Number of proposals	Funding requested (EUR)
Proposals received	54	2 905 965.00
Eligible proposals	33	1 854 529.00
Selected proposals	8	470 965.72

Table 2: Number of proposals and funding requested.

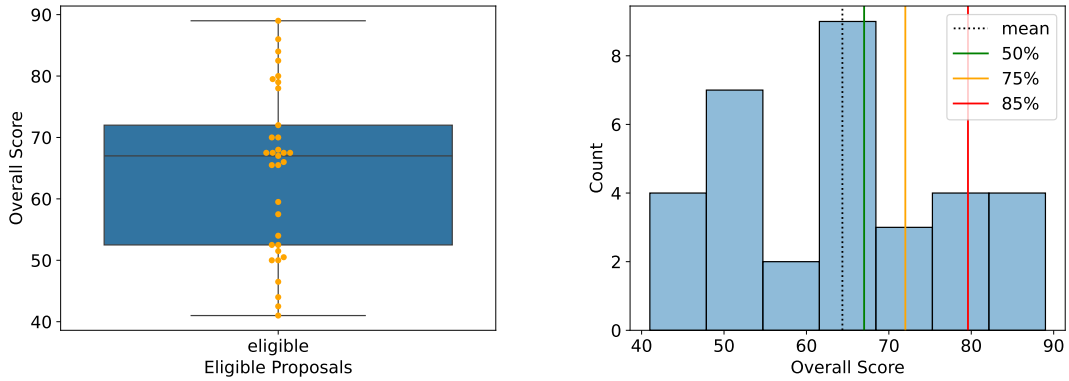


Figure 4: Distribution of Overall Score for Eligible proposals. The top cluster of 8 proposals were selected for funding.

Notification of decision. Applicants were notified of our decisions by email (sent from our HotCRP instance) on December 20, 2023 and published on UTTER's website on December 21, 2023. Besides the decisions, applicants received the complete feedback gathered throughout the evaluation procedure.

Complaints. We received one complaint via email concerning a desk-rejected submission. The applicant was based in the UK, which at the time was not considered an eligible country. Even though we had already consulted with the PO about this condition, prior to the begining of the evaluation procedure (on August 21, 2023), we did consult our PO (on January 8, 2024) to address this complaint. The PO indeed confirmed the decision was sound, given the call documentation. This was communicated to the applicant, and the complaint was settled.

1.8 Execution

Each project was assigned a 'Sponsor', a contact person within UTTER, who oversees the project execution. At a minimum, the sponsor met with the project team 3 times: at kickoff, halfway through the project duration, and at the end. In preparation for the midterm and final meetings, the project team shared a progress report covering the following:[18]

- Project execution:
  - Deviations from plan
  - Development
  - Dissemination
  - Ethics
- Summary of results and plans
  - Results

[18]Template available at https://raw.githubusercontent.com/utter-project/fstp/main/2023/UTTER_FSTP1_Report_Template.pdf.

- Business plan
- Future plans

The sponsor assesses the performance of the project team with respect to the proposed plan.

Projects started in January 2024 and ended by September 30, 2024. They received the funding in two instalments, one at the beginning and another at the end, the latter being conditioned on a positive assessment from the Sponsor.

1.9 Projects and Results

Next, we introduce the projects, a summary of their key results and a qualitative remark from the Sponsor based on the project's final report and performance. The complete reports (attached to Appendix B) contain much more detailed information on project execution, results and plans.

1.9.1 MaLA - Massive Language Adaptation of LLMs

- Recipient: University of Helsinki
- Country: Finland
- Project duration: 9 months
- Funding Awarded (EUR): 56 273.00
- Website: https://huggingface.co/MaLA-LM/emma-500-llama2-7b

Project Description. This project explores language model adaptation across multiple languages and domains to improve human-machine interaction, especially for underrepresented languages. It aims to expand language models' capabilities by collecting and fusing data in over 500 languages in various domains, addressing challenges of language diversity. It delves into continual learning methods and adaptation techniques based on existing successful model architectures and open models, increasing the accessibility and applicability of large language models, particularly for low-resource languages.

Summary of Results. The UTTER FSTP has made significant strides in advancing multilingual language models with the creation of the MaLA corpus[19] and the development of the EMMA-500 model[20]. The MaLA corpus is a diverse dataset encompassing 939 languages, 546 of which were used to train EMMA-500, a cutting-edge multilingual model. EMMA-500 has demonstrated improved performance on various language tasks such as machine translation, commonsense reasoning, and text classification across multiple languages, including low-resource languages. A pre-print is available Ji et al. (2024).

[19] https://huggingface.co/collections/MaLA-LM/mala-corpus-66e05127641a51de34d39529
[20] https://huggingface.co/MaLA-LM/emma-500-llama2-7b

Recommendation by Sponsor. The goals of this project were to collect a massively multilingual corpus and use this to train an LLM supporting a large number of languages. This has been achieved, the MaLA corpus was released and the EMMA model created by fine-tuning Llama 2 7B on this corpus. The evaluation results show strong performance, especially in MT. The data and model have both been made publicly available and there is a preprint describing them on Arxiv.

1.9.2 PenGUIn - Prototype of an ExteNded reality Graphical User INterface

- Recipient: RE:LAB Srl
- Country: Italy
- Project duration: 9 months
- Funding Awarded (EUR): 57 395.00
- Website: https://www.re-lab.it/projects/penguin

Project Description. The aim of PenGUIn is to support user experience through an innovative, inclusive, adaptive and usable Graphical User Interface for XR platforms, and study the most appropriate information design framework to support agent tasks and the relative cognitive load in the presented UTTER's use cases - and beyond (e.g., virtual learning, virtual healthcare), to support the achievement of the task objectives. The proposed solution will converge innovativeness, usability and content design in a dynamic of functionality, effectiveness, and ergonomics, according to RE:LAB's methodology "Interaction Engineering". PenGUIn aims to design and test a user-centred design approach to prototype a library of graphical elements that are intuitive, cross-cutting, and compatible with usability criteria.

Summary of Results. The project, spanning 9 months, aimed to enhance user experience through an intuitive, inclusive, and adaptive Graphical User Interface (GUI) for online platforms. This was done by studying the most appropriate information design framework and applying suitable interaction strategies to support user's tasks in the context of two case studies: a customer assistant platform and an online meeting platform. PenGUIn's concept was driven by innovation and usability to achieve functionality, effectiveness, and ergonomic experience, building on RE:LAB's user-centric methodology, "Interaction Engineering". The purpose of PenGUIn's design effort was to guide the user through the multiple platforms' functionalities, from the multilingual translation to the AI-assistant. PenGUIn UI supported transparent and task-oriented dialogue and interaction between users of these virtual platforms. The project focused on customization flexibility, going through several design iterations, and validating the prototypes through expert analysis, focus group, and testing. The work carried out in the project has represented an additional opportunity to experiment RE:LAB original proposition and new research purposes, to consolidate the team expertise in creating and testing novel user experiences. The final prototypes are available as interactive demos: Customer Assistant Interface Prototype and Meeting Assistant Interface Prototype.

Recommendation by Sponsor. The project proposal was aiming at "design and test a user-centred design approach to prototype a library of graphical elements that are intuitive, cross-cutting, and compatible with usability criteria" considering the UTTER use cases. The work

resulted in two user interface prototypes that have been tested in focus groups to evaluate their usability. These user interfaces were made available as Figma templates that could be used as a base for developing graphical user interfaces using any desired front-end framework. The project delivered on what has been proposed. To the best of our knowledge the project has been disseminated on social media channels and in the company's webpage. The project team documented business plans and possible future works including possible opportunities collaboration with one of the institutions that belong to UTTER (NAVER). Based on this, the recommendation is to approve the final payment to the project Awardee.

1.9.3 HR-XR-XTEND - Croatian XR Extensions

- Recipient: University of Zagreb, Faculty of Humanities and Social Sciences
- Country: Croatia
- Project duration: 9 months
- Funding Awarded (EUR): 60 000.00
- Website: https://hr-xr-xtend.ffzg.unizg.hr

Project Description. The project is to develop a large language model (LLM) for the Croatian language and it will be trained on a massive dataset of Croatian text. The project is aligned with the objectives of the call, as it aims to build resources for XR models, extend XR models to new language, and evaluate XR models. The project goals are to collect at least 6 billion tokens of Croatian text and prepare that data for LLM training, create a LLM for the Croatian language using monolingual data only, and evaluate the LLM for downstream tasks. The experimental phase will focus on developing and evaluating the model architecture and training process. The integration phase will involve integrating the LLM into the UTTER platform. The dissemination phase will involve disseminating the project results to the research community and the public.

Summary of Results. The project aimed to create a large-scale monolingual Croatian language model (HR-GPT Beta). A significant training dataset was collected and cleaned from existing mono- and multilingual resources that include texts in Croatian. The preprocessing featured also advanced deduplication techniques, resulting in a final training dataset of 7.72 billion tokens. Three training scenarios were used: training from scratch, continued pretraining on a monolingual model, and continued pretraining on a multilingual model. The evaluation was performed using several benchmark datasets, and fine-tuning with the Alpaca dataset improved model performance. Larger models, like "gemma-7b", outperformed smaller ones, and fine-tuning enhanced results further. Key results include multiple model versions (160M, 350M, 410M, and 1.4B parameters) and a cleaned training dataset. Future work involves additional data collection, additional model training, further NLP task evaluations, and more training experiments. The HR-GPT Beta and training material (partially) will be publicly accessible under permissive licenses from the HR-CLARIN repository (https://clarin.hr). More information can be found on the project website.[21]

[21] https://hr-xr-xtend.ffzg.unizg.hr

Recommendation by Sponsor. The project planned the collection of Croatian datasets and training a large language model on Croatian and they delivered on this objective. This project was successfully disseminated. UTTER could use these datasets for training the EuroLLM language model. This project has completed successfully.

1.9.4 SignReality - Extended Reality for Sign Language Translation

- Recipient: DFKI GmbH
- Country: Germany
- Project duration: 9 months
- Funding Awarded (EUR): 59 995.22
- Website: https://www.dfki.de/en/web/research/projects-and-publications/project/signreality

Project Description. SignReality will create a 3D sign language interpreter displayed in Augmented Reality glasses. It will serve as an extension of the UTTER online/hybrid interfaces, aiming at usability and accessibility for deaf and hard-of-hearing people. The app will be based on an XR model consisting of a pre-trained sequence-to- sequence neural network, connected to a framework for geometrical transformations for synthesizing an animated avatar. This will follow a client-server architecture, connected with the SDK of the AR device and via an API to other apps. Participatory design and evaluation in co-operation with the user community is planned. Results will be disseminated to the user and scientific communities, to UTTER and parallel research projects and will be used to initiate further research.

Summary of Results. The project achieved significant milestones in bridging sign language technology with Extended Reality. Key results include the development of an engine for avatar animation, accompanied by device-specific implementation on two AR devices (Hololens 2 and XReal Light). Translation from spoken language to a textual sign language representation (German → DGS) was enabled through an encoder-decoder translation model, whereas further improvement of relevant models will benefit from the work on corpus acquisition and alignment. The implementation was tested for intelligibility, user experience and acceptance in a user study with native sign language users at the Centre for Culture and Visual Communication of the Deaf Berlin Brandenburg (ZFK) providing valuable feedback. The project has been integrated into several academic theses and university workshops, and research findings will be submitted in relevant academic venues.

Recommendation by Sponsor. The project clearly achieved all of its goals, with very minor deviations from the original plan, both along the scientific and dissemination dimensions. The project team has experience with the ethical considerations behind the experimental setup and did a remarkable job both at complying with all relevant guidelines and regulations but also at clearly documenting the scope of their findings and technology.

1.9.5 DeMINT - Automated Language Debriefing for English Learners via AI Chatbot Analysis of Meeting Transcripts

- Recipient: University of Alicante
- Country: Spain
- Project duration: 9 months
- Funding Awarded (EUR): 57 567.50
- Website: https://github.com/transducens/demint

Project Description. This project focuses on developing an AI chatbot to serve as a tutoring assistant for non-native English speakers, enhancing their language skills through post-meeting analysis of meeting transcriptions. This effort aligns with UTTER's objectives, particularly its interest in harnessing language models for video conferencing applications. Following recent advances in LLM-based chatbots and agents, our system will exploit pre-trained large language models, refined for the tutoring task through a mix of in-context learning, external non-parametric memory retrieval, efficient parameter fine-tuning, and tool exploitation. Human evaluation will be conducted through individual debriefings after simulated, scenario-based video conferences with small test groups.

Summary of Results. DeMINT has developed a prototype of a conversational system designed to enhance non-native English speakers' language skills through post-meeting analysis of the transcriptions of video conferences in which they have participated. The code of the system is already available as open-source software on https://github.com/transducens/demint, and a paper (Pérez-Ortiz et al., 2024) has been published at the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning. Future plans include developing a more engaging and speech-based interaction with the chatbot and knowledge from theories of second language acquisition.

Recommendation by Sponsor. The DeMINT team has done an excellent job at delivering all results promised in the project proposal. A publication at a relevant workshop and an open-source codebase for the tool developed have been disseminated. The project has not deviated in any major way from what was proposed. The project sponsor, therefore, makes a positive payment recommendation for the DeMINT project.

1.9.6 SURE-GB - Identifying under-representational, stereotypical and algorithmic gender bias in machine translation

- Recipient: Institute of Computer and Communicati on Systems, National Technical University of Athens, ICCS-NTUA
- Country: Greece
- Project duration: 9 months

- Funding Awarded (EUR): 60 000.00
- Website: https://ailswp.ails.ece.ntua.gr/suregb/

Project Description. SURE-GB aims to build an automated service that identifies occupation-related under representational, stereotypical, and algorithmic gender bias in machine translation, in English and French, as well as low resource languages like Greek. The proposed method involves creating a curated knowledge graph that a) encodes standardised knowledge and data for occupations (based on data and hierarchies from EU- LFS1, the ESS2, and the International Classification of Occupations-ISCO3), b) incorporates statistics for occupation-related gendered language usage derived from linguistic corpora. Our goal is to develop a ready-to-use machine learning toolkit, that utilises the above knowledge to detect and categorise gender biases for: a) providing actionable recommendations for improvement, b) establishing guidelines for unbiased language translation, c) raising awareness of gender biases in machine translation systems.

Summary of Results. The SURE-GB project has made significant strides in understanding and addressing occupational gender biases in machine translation (MT) systems. Our research has resulted in the creation of a curated Knowledge Graph that encapsulates essential statistics on gender representation in various occupations across Greece, France, and the UK, as well as linguistic biases in corresponding textual corpora. We have developed an automated system to detect and classify occupational gender bias in MT systems. By revealing the disparities between real-world statistics and their representation in MT systems, we have identified critical flaws in current technologies and paved the way for more equitable and accurate translations. Through this project, we enable a more nuanced study of machine learning biases by disentangling the real world, the data, and the systems, while still recognizing their interconnectedness. Explore our findings through our website.

Recommendation by Sponsor. Given the clear achievements, adherence to the proposal, and the potential for future impact and expansion, the SURE-GB team has shown more than satisfactory performance and already produced several outputs. Hence, it is recommended that the SURE-GB project receives the final funding part as originally planned.

1.9.7 InCroMin - Interactive Crosslingual Minutes

- Recipient: Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
- Country: Czech Republic
- Project duration: 9 months
- Funding Awarded (EUR): 59 826.25
- Website: https://github.com/ELITR/incromin-test-calls

Project Description. The goal of the project is to (1) expand existing interactive meeting summarization tools (such as our MinuteMan, or those developed by UTTER) to facilitate cross-lingual access to meeting content (live transcripts and live minutes) and (2) make these tools benefit from human interpretation, if available in the meeting. As a necessary prerequisite, the project will prepare a test set and rigorously evaluate the underlying models of speech transcription, translation and summarization in this setting.

Summary of Results. InCroMin examined and carefully evaluated the applicability of recent state-of-the-art speech-to-text translation tools in real cross-lingual calls, i.e. calls between parties that do not have a common language. The project adapted MinuteMan (https://github.com/fkmjec/minuteman) for this purpose and collected a corpus of such calls. The deidentified part of the corpus is available here: https://github.com/ELITR/incromin-test-calls. Additional results of InCroMin include an evaluation of latency metrics for speech translation, translation of ELITR-Bench (https://github.com/utter-project/ELITR-Bench) into Czech to allow evaluation of cross-lingual access to past meeting content or translation of a part of MultiWOZ dialogues into Czech and German to assess translation quality of dialog-critical features such as participants' gender preservation. All the outputs are detailed in InCroMin Final Report.

Recommendation by Sponsor. The project exceeded expectations. In summary, they i) adapted MinuteMan to support cross-lingual calls, ii) collected a new and potentially valuable corpus of simulated cross-lingual meetings, and iii) conducted practical tests to assess the usability of the extended MinuteMan and identified areas for improvement. For well-supported languages, MinuteMan appears close to being fully operational. Additionally, the potential founding of a spin-off for MinuteMan is under consideration, with FSTP funding playing a crucial role in bringing the system closer to production-ready. Finally, even though it wasn't initially planned, InCroMin developed a Czech version of the ELITR-Bench meeting, which will soon be added to the UTTER/ELITR-Bench repository. This could also spark future collaboration between Naver and Charles University on cross-lingual QA on long documents (meeting transcripts).

1.9.8 pyannote.mobile

- Recipient: Université Toulouse III – Paul Sabatier
- Country: France
- Project duration: 9 months
- Funding Awarded (EUR): 59 908.75
- Website: https://pyannote.ai

Project Description. pyannote.mobile aims at extending pyannote speaker diarization open-source toolkit in two complementary directions. The first one is to add streaming speaker diarization support, as it currently only supports offline/batch processing. The second one is to investigate the feasibility of "on device" streaming speaker diarization (as opposed to cloud-based processing): we will develop a streaming speaker diarization proof-of-concept running on mobile

(iOS or Android). For both directions, we will aim for the best compromise between accuracy and (algorithmic and computational) latency.

Summary of Results. pyannote.mobile project led to the extension of the pyannote.audio open-source speaker di- arization toolkit to perform speaker diarization in real-time while controlling the trade-off between latency and accuracy. It also led to the creation of an iOS/macOS streaming speaker diarization SDK which will be handed over to interested parties through the local university tech transfer office.

Recommendation by Sponsor. As the sponsor of this project, we confirm that the project successfully delivered its planned results. The dissemination efforts were effective, including an iOS application soon to be available, and a scientific paper published at Interspeech 2024. The project lead also provided comprehensive documentation and effective communication throughout the duration of the project, which were appropriate and well-aligned with the project's objectives. Overall, we recommend this project positively, as it has met its key objectives and demonstrated potential for future impact.

1.10 Exploitation

It is relatively early for further exploitation of the results, esp. within UTTER itself, as the projects have just recently concluded. We expect to have a better overview of this towards M36 (when a follow-up deliverable is due, to report on our 2nd FSTP call and ourFSTP programme as a whole).

On July 5, UTTER hosted its 2nd User Day, an online event which, amongst other things, advertised our 2nd FSTP call. At that event, some of the FSTP1 awardees (namely, DeMINT, SURF-GB and InCroMin) joined to present their progress and demonstrate to event attendees (potential applicants for our 2nd FSTP call) the impact of our FSTP programme.

Each project is engaging with their own plans for exploitation, in some cases this has already led to collaborations with UTTER, e.g. (Ji et al., 2024), in other cases a closer collaboration with UTTER is envisioned (e.g., in RP2), e.g., PenGUIn, HR-XR-XTEND, InCroMin.

2 Conclusion

Our first FSTP call was by all means a success. It attracted many submissions, many of which were of high-quality, and 8 of which were selected for funding. The awardees executed their projects on time and successfully, they disseminated their projects adequately and their project's outputs are in almost every instance directly available to the public. In a few instances, their outputs are likely to impact UTTER within its final year (e.g., curated data that can be used by our models, or UI design and ideas that may affect our final prototypes). They indicate plans for the future ranging from extended research papers, to connecting to industry and/or communities of users, to securing additional funding. These 8 projects constitute an important part of UTTER's impact, and we will be keeping a close eye on how they develop beyond the 9 months in which they were funded by our FSTP programme.

References

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. Emma-500: Enhancing massively multilingual adaptation of large language models, 2024. URL https://arxiv.org/abs/2409.17892.

Juan Antonio Pérez-Ortiz, Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Felipe Sánchez-Martínez, Roman Chernysh, Gabriel Mora-Rodríguez, and Lev Berezhnoy. A conversational intelligent tutoring system for improving English proficiency of non-native speakers via debriefing of online meeting transcriptions. In Thomas Gaillat, Cyriel Mallart, Fabienne Moreau, Jen-Yu Li, Griselda Drouet, David Alfter, Elena Volodina, and Arne Jönsson, editors, *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 187–198, Rennes, France, October 2024. LiU Electronic Press. URL https://aclanthology.org/2024.nlp4call-1.14.

A Screenshots of Review Forms

Free of CoI *

I have no conflict of interest with this submission.

Formal requirements - checklist *

Formal requirements check is the first step in the evaluation process. If one of the formal requirements is not fulfilled, the proposal is rejected.

1. Language: Proposal is in English in all required parts.
2. Submission: Proposal delivered on time, through the designated system, using the requested template.
3. Declaration of Honour: Declaration of Honour is signed.
4. Legal Status: Applicant is an SME or research organisation (incl., but not limited to, higher education organisations, independent research organisations and NGOs).
5. Country: Applicant is legally established in a Horizon Europe eligible country.
6. Number of Proposals: Maximum of one proposal per applicant.
7. Conflict of Interest: No conflict of interest.
8. Complete: All required sections of the proposal are filled in.

Comments on formal requirement

Use this field to make remarks about violated formal requirements and missing information.

Amendments

You can list amendments that must be done (or at least considered) in case the proposal goes on to the **contract signing** phase.

Formal requirements - recommendation *

We desk-reject submissions that fail to comply with the criteria in the formal requirement checklist. If a proposal complies with those criteria, but requests more budget than the maximum allowed in this call (i.e., €60,000 per pilot project), it may be moved to the next phase of the assessment, but, should it get to the contract signing phase, it will require amendments. Note that, the proposal may still be deemed unviable in the next assessment phase (even before any amendment is requested).

1. Desk-reject due to failure to comply with formal requirements
2. Move to next phase, but mark for amendment.
3. Move to next phase

Figure 5: Screenshot of review form for Formal Requirements.

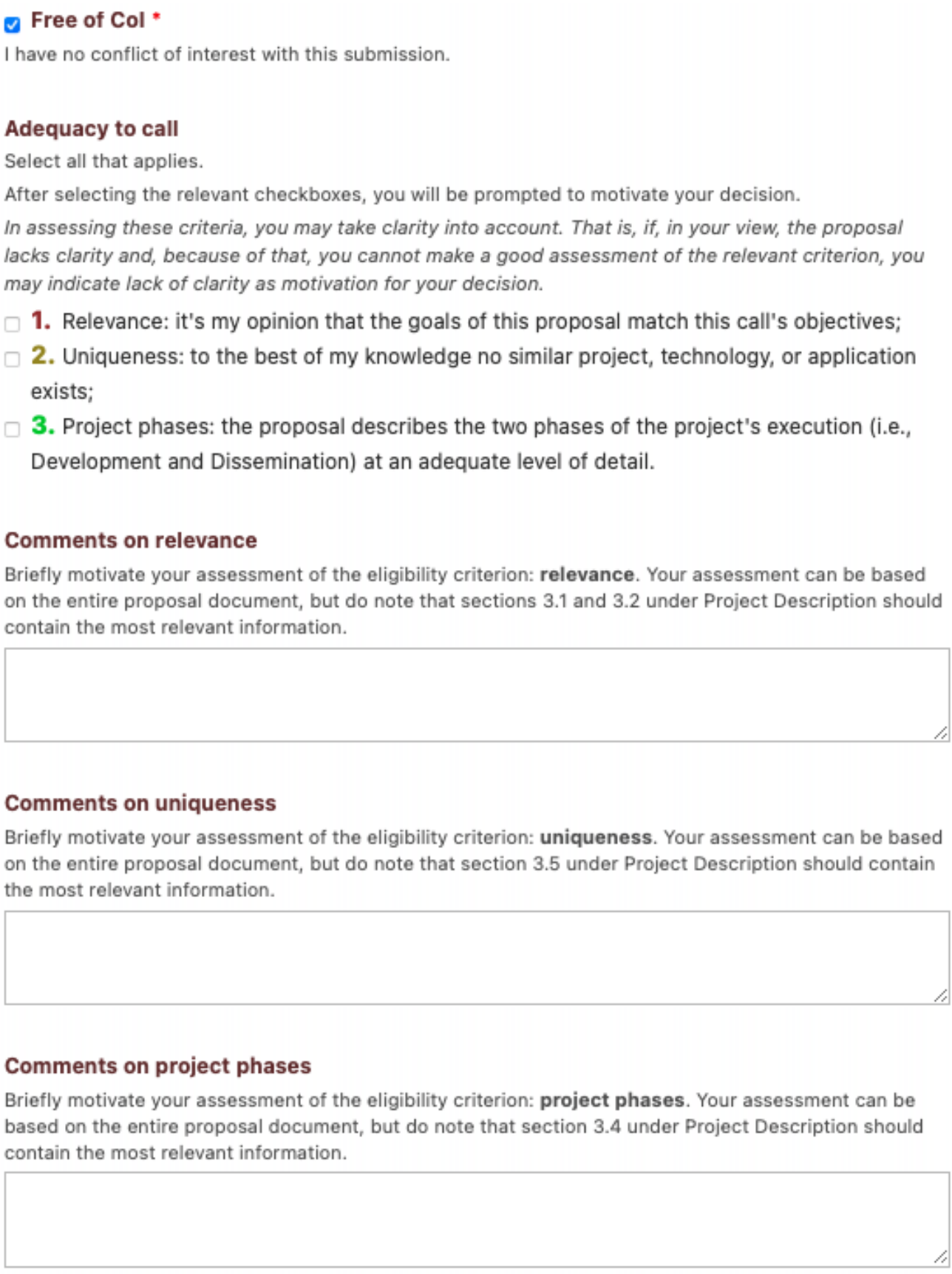


Figure 6: Screenshot of review form for Adequacy to Call (Eligibility).

Free of CoI *

I have no conflict of interest with this submission.

Key parameters *

I am familiar with the key parameters of this call:

Objectives

Develop and/or pilot applications using XR models (i.e., pre-trained neural network models adaptable to a large variety of forms of expression, interaction, languages, domains, styles and intent) in new sectors, with a focus on enabling new types of human-human and human-machine interaction. Examples of welcome project objectives include:

- Improving or demonstrating efficiency of XR model inference;
- Improving or demonstrating efficiency of XR model training;
- Designing interfaces for usability;
- Extending XR models to new languages, domains or modalities;
- Applying XR models to new tasks;
- Building resources for XR models;
- Evaluation of XR models.

Proposals

- Maximum budget per project: 60,000 euro
- Project duration: 6 months
- Applicant: SME or research organisation for a Horizon Europe eligible country

Project execution

- Development
- Dissemination

Evaluation criteria

- Objective fit
- Technical approach
- Business, Integration and Dissemination (BID) plan
- Budget adequacy
- Team
- Ethics
- Evaluation of XR models.

Criteria fulfilment

Score	Rubric
0	Not at all: The proposal fails to address the criterion or cannot be assessed due to missing or incomplete information
3	Limited: The criterion is inadequately addressed or there are significant weaknesses.
7	Good: The proposal addresses the criterion well, but some shortcomings are present.
10	Excellent: The proposal successfully addresses all relevant aspects of the criterion, any shortcomings are minor.

Figure 7: Screenshot of review form for Qualitative Assessment (1/6) - Key Parameters

Objective fit - clarity *

Are the project goals clear?

- 1. The goals aren't explicitly outlined, I had to infer what the goals might be by studying their proposed plan.
- 2. The proposal outlines its goals, but I do not understand them well (e.g., the proposal lacks detail, it is difficult to appreciate the key arguments without specialised knowledge I don't have).
- 3. The proposal clearly outlines its goals, they are explained well and I understand them.
- 4. Accept
- 5. Strong accept

Objective fit - adequacy *

Are the project goals and planned achievements in line with the overall objectives of UTTER?

- 1. The proposal treats its goals as somewhat intuitively obviously aligned with UTTER's goals for the call, and I don't agree with this assumption.
- 2. The proposal justifies how its goals are aligned with UTTER's objectives for this call, but I do not agree with some of the arguments. For example, I disagree with certain premises or predictions.
- 3. The proposal justifies how its goals are aligned with UTTER's objectives for this call, I find the justification reasonable and I agree with the arguments in the proposal.
- 4. Expert

Objective fit - impact *

Is it likely that the project will deliver added value to UTTER? There are various ways to add value. Here are some examples: i) resources (e.g., a dataset, a UI, a set of requirements), ii) methods (e.g., a technique for training, or inference), iii) results (e.g., observations about tools, users, or datasets), or iv) a position (e.g., a critical investigation of key premises, a careful outline of ethical considerations, a discussion about broader impact or implications of technology) that UTTER (or the larger body of work around UTTER) can build upon; or through v) an original demonstration of the impact that XR technology (developed by or relevant to UTTER) can have outside academia. You may also recognise some other mechanism which you believe has a similarly important value for UTTER.

- 1. I do not think the project will add value to UTTER. Potential contributions are marginally relevant.
- 2. There is potential for some added value along one or more dimensions such as (i-v) above. The contributions are, however, uninspiring and may go unnoticed, or they are unlikely to affect UTTER and the larger body of work around UTTER within UTTER's lifespan.
- 3. The project will clearly add value to UTTER along one or more dimensions such as (i-v) above, UTTER and the larger body of work around UTTER will likely benefit from it within UTTER's lifespan.

Objective fit: overall score *

Assign an overall score along the *objective fit* dimension. Base your scores on your choices for the sub-criteria *objective fit: clarity*, *objective fit: adequacy* and *objective fit: impact*. As a guideline, you can multiply the marks you assigned for the sub-criteria and find a suggested overall score range below. You can deviate from this suggestion, but major deviations might be a sign that you need to reconsider your assessment of the sub-criteria.

Product	Score range
1	1-3
2-3	2-4
4	3-5
6-9	4-6
12	5-7
18	6-8
27	8-10

(Choose one)

Comments on objective fit

You can use this box to make comments on objective fit. For example, to justify an extreme score, to explain a deviation from the rubrics, or to document a specific interpretation of a rubric, etc. Your comments will be visible to authors.

Figure 8: Screenshot of review form for Qualitative Assessment (2/6) - Objective fit

Technical approach - feasibility *

Are the planned activities feasible and facilitate the achievement of project outputs? As a reminder, this call invited proposals to develop project ideas over a period of 6-9 months with a maximum budget of 60 thousand Euro.

- **1.** The activities are documented insufficiently or the proposal lacks clarity. Or, the planned activities aren't realistic given the allocated resources. Or, there's an incongruence between the planned activities and the intended project output.
- **2.** The planned activities are reasonably aligned with the project goals, they may however be somewhat incompatible with the allocated resources.
- **3.** The activities are described clearly, they are aligned with the project outputs and compatible with the allocated resources.

Technical approach - originality *

Does the proposal push the boundaries of existing XR technology?

- **1.** The approach was discussed unclearly or at an insufficient level of detail. Or, I find the approach uninteresting, trivial or redundant.
- **2.** The approach is presented at a reasonable level of detail and I recognise some potentially original elements.
- **3.** The approach is presented at a reasonable level of detail, it contributes creatively to XR technology and/or application.

Technical approach: overall score *

Assign an overall score along the *technical approach* dimension. Base your scores on your choices for the sub-criteria *Technical approach: feasibility* and *Technical approach: originality*. As a guideline, you can multiply the marks you assigned for the sub-criteria and find a suggested overall score range below. You can deviate from this suggestion, but major deviations might be a sign that you need to reconsider your assessment of the sub-criteria.

Product	Score range
1	1-3
2	2-4
3	3-5
4	4-6
6	6-8
9	8-10

(Choose one)

Comments on technical approach

You can use this box to make comments on technical approach. For example, to justify an extreme score, to explain a deviation from the rubrics, or to document a specific interpretation of a rubric, etc. Your comments will be visible to authors.

Figure 9: Screenshot of review form for Qualitative Assessment (3/6) - Approach

BID plan: business *

Is the business plan reasonable and ambitious?

- 1. No (the business plan is unclear, missing, or unrealistic).
- 2. Somewhat (it might lack some ambition, or be a little too ambitious, or lack detail).
- 3. Yes.

BID plan: integration *

How well is the integration of project outputs planned?

- 1. Poorly. There are too many missing pieces, or it builds on non-existing resources and technology without a clear mitigation strategy, etc.
- 2. Decently, but I would have appreciated more detail, or I doubt the feasibility of some aspects and the plan did not discuss any contingencies.
- 3. Good plan presented at a good level of detail including contingencies where needed.

BID plan: dissemination *

Are the dissemination and promotion activities planned adequately?

- 1. No. The strategies aren't effective or too vague.
- 2. Reasonably well, it will probably reach the relevant target audiences.
- 3. Remarkably well, it's clear who the target audiences are and how they will be approached.

BID plan: overall score *

Assign an overall score along the *BID plan* dimension. Base your scores on your choices for the sub-criteria *BID plan: business*, *BID plan: integration* and *BID plan: dissemination*. As a guideline, you can multiply the marks you assigned for the sub-criteria and find a suggested overall score range below. You can deviate from this suggestion, but major deviations might be a sign that you need to reconsider your assessment of the sub-criteria.

Product Score	range Rubric
1	1-3
2-3	2-4
4	3-5
6-9	4-6
12	5-7
18	6-8
27	8-10

(Choose one)

Comments on business, implementation and dissemination plan

You can use this box to make comments on BID plan. For example, to justify an extreme score, to explain a deviation from the rubrics, or to document a specific interpretation of a rubric, etc. Your comments will be visible to authors.

Figure 10: Screenshot of review form for Qualitative Assessment (4/6) - BID

Team *

Is the applicant's team capable of executing the project and delivering its outputs (in required time, quality and with estimated budget)?

Score	Rubric
0	Not at all: The proposal fails to address the criterion or cannot be assessed due to missing or incomplete information
3	Limited: The criterion is inadequately addressed or there are significant weaknesses.
7	Good: The proposal addresses the criterion well, but some shortcomings are present.
10	Excellent: The proposal successfully addresses all relevant aspects of the criterion, any shortcomings are minor.

(Choose one)

Comments on team

You can use this box to make comments on team. Your comments will be visible to authors.

Budget *

Does the budget correspond to all panned activities and outputs?

Score	Rubric
0	Not at all: The proposal fails to address the criterion or cannot be assessed due to missing or incomplete information
3	Limited: The criterion is inadequately addressed or there are significant weaknesses.
7	Good: The proposal addresses the criterion well, but some shortcomings are present.
10	Excellent: The proposal successfully addresses all relevant aspects of the criterion, any shortcomings are minor.

(Choose one)

Comments on budget

You can use this box to make recommendations regarding the budget. Your comments will be visible to authors.

Figure 11: Screenshot of review form for Qualitative Assessment (5/6) - Team and Budget

Ethics *

Is the ethical self-assessment thoughtful and thorough? Does it provide convincing justification that the applicant will ensure the work will be done ethically?

Score	Rubric
0	Not at all: The proposal fails to address the criterion or cannot be assessed due to missing or incomplete information
3	Limited: The criterion is inadequately addressed or there are significant weaknesses.
7	Good: The proposal addresses the criterion well, but some shortcomings are present.
10	Excellent: The proposal successfully addresses all relevant aspects of the criterion, any shortcomings are minor.

(Choose one)

Comments on ethics

You can use this box to make remarks and/or recommendations regarding the ethics self-assessment. For example, if you don't think the self-assessment is thorough, you can highlight and defend this here. If you believe the assessment is thorough but the mitigation strategies aren't adequate, do highlight this here. Use this space for any other advice and/or recommendation. Your comments will be visible to authors.

Comments for PC (hidden from authors)

Figure 12: Screenshot of review form for Qualitative Assessment (6/6) - Ethics and Comment for PC

B Reports from Project Teams

B.1 MaLA

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP1 Final – MaLA

Massive Language Adaptation

Nature	Final Report	**Work Package**	WP1
Project start date	15/01/2024	**Project end date**	07/10/2024
Interim meeting	28/05/2024	**Report submission Date**	07/10/2024
Main authors	Barry Haddow (UEDIN)		
Co-authors	Shaoxiong Ji (University of Helsinki)		
Reviewers	Maryam Hashemi (UVA)		
Version Control			
v0.1	**Status**	Draft	XXX
v1.0	**Status**	Final	XXX

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Project Execution

1.1 Deviations from original plan

There are no significant changes or deviations from orginal plan.

1.2 Development

Compilation of the MaLA corpus

The MaLA, **Ma**ssive **L**anguage **A**daptation, corpus has been successfully compiled, containing data from 939 languages. Of these, 546 languages with over 100,000 tokens have been selected for training the EMMA-500 model. The corpus offers diverse data types, including code, books, scientific papers, and instruction data, with more than 100 billion whitespace-delimited tokens. Four versions of the corpus have been made available, meeting different processing needs: noisy, cleaned, deduplicated, and split versions.

Extension of the MaLA corpus

The MaLA corpus[1] has been extended by integrating multiple curated datasets. This augmentation resulted in a rich, diverse data mix that supports the continual pre-training of large language models, ensuring a comprehensive dataset for enhanced language adaptation across a broad range of linguistic contexts.

Continual pre-training of the EMMA-500 model

The continual pre-training of the EMMA-500 model[2] has been completed using the Llama 2 7B model (Touvron et al., 2023). The training involved 546 languages and a massive multilingual corpus, leading to the development of a model that has been rigorously evaluated across various tasks.

Evaluation and benchmarking

The EMMA-500 model has been evaluated against other multilingual and decoder-only LLMs on a wide range of tasks, including commonsense reasoning, machine translation, open-ended generation, text classification, and natural language inference. We also composed a novel multilingual benchmark, called PolyWrite[3] in this work for evaluating open-ended generation in 240 languages.

1.3 Dissemination

We release a preprint on arXiv (Ji et al., 2024), a model and different versions of the datasets on Huggingface. We disseminated these in Helsinki-NLP twitter[4], and received 17 retweet, 65 likes,

[1] https://huggingface.co/collections/MaLA-LM/mala-corpus-66e05127641a51de34d39529
[2] https://huggingface.co/MaLA-LM/emma-500-llama2-7b
[3] https://huggingface.co/datasets/MaLA-LM/PolyWrite
[4] https://x.com/HelsinkiNLP/status/1840669891101172149

and 3.1k views as of Oct 3, 2024.

1.4 Ethics

Multilingual language models trained on large, diverse datasets risk inheriting and amplifying societal biases present in the data. There is a danger that these biases might manifest in harmful ways, particularly in sensitive applications such as content generation, machine translation, and customer support. Despite our focus on low-resource languages, there remains a risk that certain languages or dialects may still be underrepresented, leading to poorer performance or exclusion from language technology advancements.

To mitigate the risk of underrepresentation, the MaLA corpus is continuously expanded to include more data from low-resource languages. By employing continual pre-training and bilingual datasets, we enhance the model's performance on these languages. We also emphasise that our released model is a "foundation model", intended for research purposes, in that it has not undergone the extensive safety testing that would be required for a deployed model. In fact, safety testing of LLMs for diverse languages is an active area of research.

2 Summary of Results and Plans

2.1 Results

In a comparison with decoder-only LLMs, including Llama 2-based continual pre-trained models and LLMs that are designed to be multilingual, the EMMA-500 model has outperformed several baselines, including Llama 2-based models, in most tasks, and outperformed some strong baselines in some tasks, showing strong progress toward improved multilingual performance. Our model achieves strong results:

- Out of models with parameter sizes from 4.5B to 13B, our model with 7B parameters has the lowest negative log-likelihood according to an intrinsic evaluation.
- Our model remarkably improves the performance of commonsense reasoning, machine translation, and open-ended generation over Llama 2-based models and multilingual baselines, and outperforms the latest advanced models in many cases.
- Our model improves the performance of text classification and natural language inference, outperforming all Llama 2-based models and LLMs designed to be multilingual.
- While math and machine reading comprehension (MRC) tasks are challenging for the Llama 2 7B model and other multilingual LLMs, our model remarkably enhances the Llama 2 base model. Our model yields improved performance on MRC over the base model but still produces quasi-random results similar to other multilingual baselines.
- We demonstrate that massively multilingual continued pre-training does not necessarily lead to regressions in other areas, such as code generation, if the data mix is carefully curated. Our model surpasses the Llama 2 7B base model's code generation abilities.

2.2 Business plan

We will need to apply for funding to support our future plans in Section 2.3.

2.3 Future plans

The next phase of this project will focus on expanding the training data and improving model evaluation processes. The planned steps are as follows:

1. **Preparation of Data Mix #2**: We are in the process of preparing Data Mix #2, which will include the content from Mix #1 used in training the released model along with additional bilingual texts and extra code/reasoning data. This dataset will contain approximately 300 billion tokens in total and will be used for the continued training of the Llama 3 model (Dubey et al., 2024).
2. **Preparation of Data Mix #3**: We also plan to work on preparing Data Mix #3, which will primarily feature more bilingual texts. The total token count for this dataset is to be determined, and the data will be used for training the Llama 3.1 model.
3. **Implementation of Additional Evaluation Tasks**: We will implement evaluation codes for two additional machine translation (MT) benchmarks, namely NTREX and Flores+, as well as an additional task of token classification. These evaluations will further assess the performance of our multilingual models across different linguistic and computational tasks.

These next steps aim to significantly advance the multilingual capabilities of our models, further improving the representation of low-resource languages and enabling better cross-lingual transfer in tasks such as machine translation and token classification.

2.4 Blurb for public dissemination on UTTER's website

The UTTER FSTP has made significant strides in advancing multilingual language models with the creation of the MaLA corpus[5] and the development of the EMMA-500 model[6]. The MaLA corpus is a diverse dataset encompassing 939 languages, 546 of which were used to train EMMA-500, a cutting-edge multilingual model. EMMA-500 has demonstrated improved performance on various language tasks such as machine translation, commonsense reasoning, and text classification across multiple languages, including low-resource languages.

3 Recommendation by Project Sponsor

The goals of this project were to collect a massively multilingual corpus and use this to train an LLM supporting a large number of languages. This has been achieved, the MaLA corpus was released and the EMMA model created by fine-tuning Llama 2 7B on this corpus. The evaluation results show strong performance, especially in MT. The data and model have both been made publicly available and there is a preprint describing them on Arxiv.

[5] https://huggingface.co/collections/MaLA-LM/mala-corpus-66e05127641a51de34d39529
[6] https://huggingface.co/MaLA-LM/emma-500-llama2-7b

References

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. EMMA-500: Enhancing massively multilingual adaptation of large language models. *arXiv preprint 2409.17892*, 2024. URL https://arxiv.org/abs/2409.17892.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*, 2023.

B.2 PenGUIn

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP1 Final – PenGUIn

PenGUIn

Nature	Final Report	**Work Package**	WP1
Project start date	09/02/2024	**Project end date**	30/09/2024
Interim meeting	19/06/2024	**Report submission Date**	15/10/2024
Main authors	José Souza (UNB), Pedro Martins (UNB)		
Co-authors	Stefania Aguzzi (RE:Lab)		
Reviewers	Maryam Hashemi (UVA)		
Version Control			
v0.1	**Status**	Draft	09/10/2024
v1.0	**Status**	Final	09/10/2024

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Project Execution

This is the Final report of the PenGUIn, presenting the progress made since the previous report delivered in June 2024 and later presented to the UTTER team at the interim meeting. The aim of the project is to support UTTER by designing an innovative and user-friendly Graphical User Interface for the two use cases (UCs), the online customer service assistant and the meeting assistant. The work carried out by PenGUIn will, on one side, improve the usability and the user experience, while enhancing, on the other, the functionalities of the UTTER platforms. In this later phase, from June to September, RE:LAB has focused on implementing the feedback received from UTTER about the latest changes to the prototypes, leading to their finalization. In terms of dissemination, the final prototypes were also published on RE:LAB social media channels.

1.1 Deviations from original plan

No deviations occurred.

1.2 Development

Development of the Use Case 1 - Customer Service Assistant

Our work here was dedicated to the implementation of a colour code scale to visualize the range of values of the COMET index. The colour code has replaced the values from the previous versions of the prototypes to be clearer and intuitive. In fact, this was one of the feedback items emerging from the focus groups with experts held in June. Figure 1 shows the last version of the interface for this use case.

Figure 2 shows how the colour code works, with the interface providing a colour line under each chat box corresponding to the specific scale value.

Development of the Use Case 2 - Meeting Assistant

The updates included the implementation of the light mode of the interface and the reduction of the number of meeting lines listed in the dashboard. As per the previous use case, they reflected the comments shared during the focus group. The figures below show the latest modifications (Figure 3 and Figure 4).

1.3 Dissemination

RE:LAB has implemented a series of promotional activities aiming to disseminate and give visibility to the project. Examples of the promotion activities include:

- Project website in the company website
- Article dedicated to the project published on the company website
- Post on RE:LAB LinkedIn page
- Post on RE:LAB Instagram page

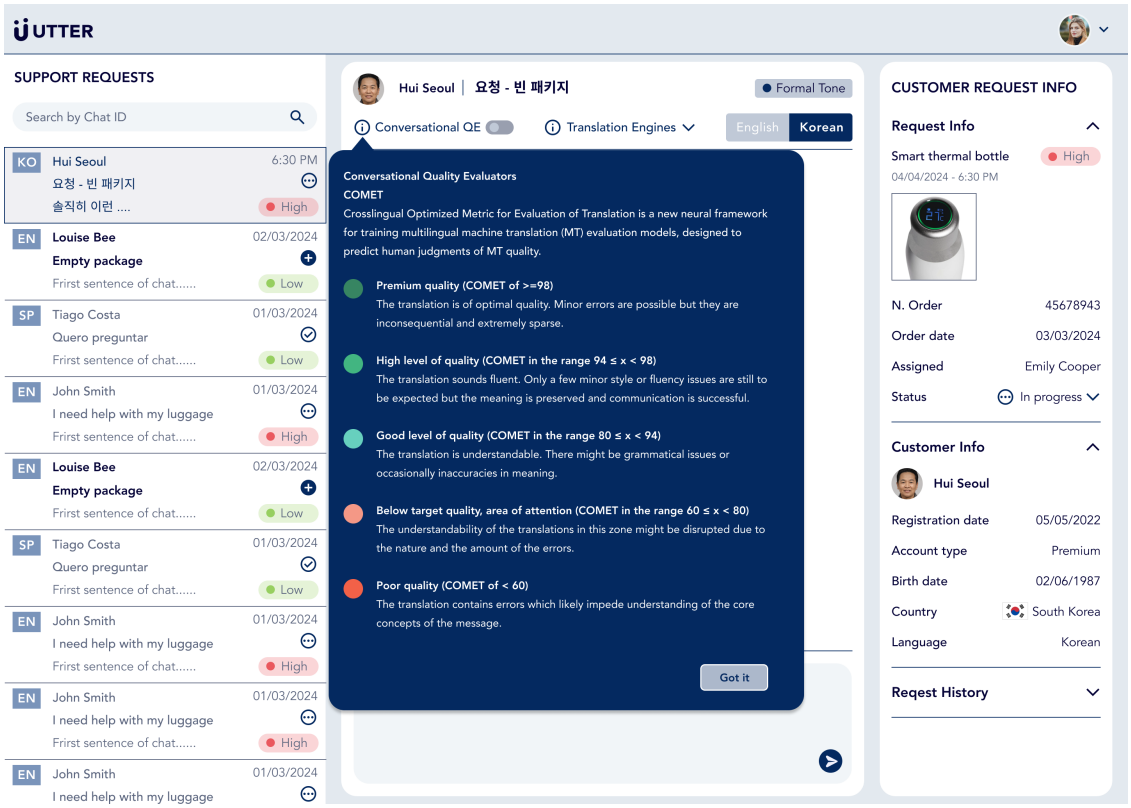


Figure 1: Customer Service Assistant – Last version.

1.4 Ethics

Although the ethical implications of the activities carried out in this project were limited due to their nature, RE:LAB applied standard compliance and prevention strategies to ensure the protection of personal data as well as the monitoring of potential risks.

In terms of project data management, the project has not produced relevant or sensitive data itself. All project materials have been stored in a company's secure virtual environment that is open only to the project team.

As far as it concerns the prototyping of the two interfaces, no particular requirements of privacy/ data protection were preliminarily identified nor raised, so the design activities have followed the standards applicable to these kinds of platforms.

With regards to the organisation of the focus group with UX/UI experts, RE:LAB applied the standard data protection strategies for the event. This was done by informing the participants about how their personal information would have been used and asking for consent to share information and event materials. Consent was also asked in order to recorder the online meeting.

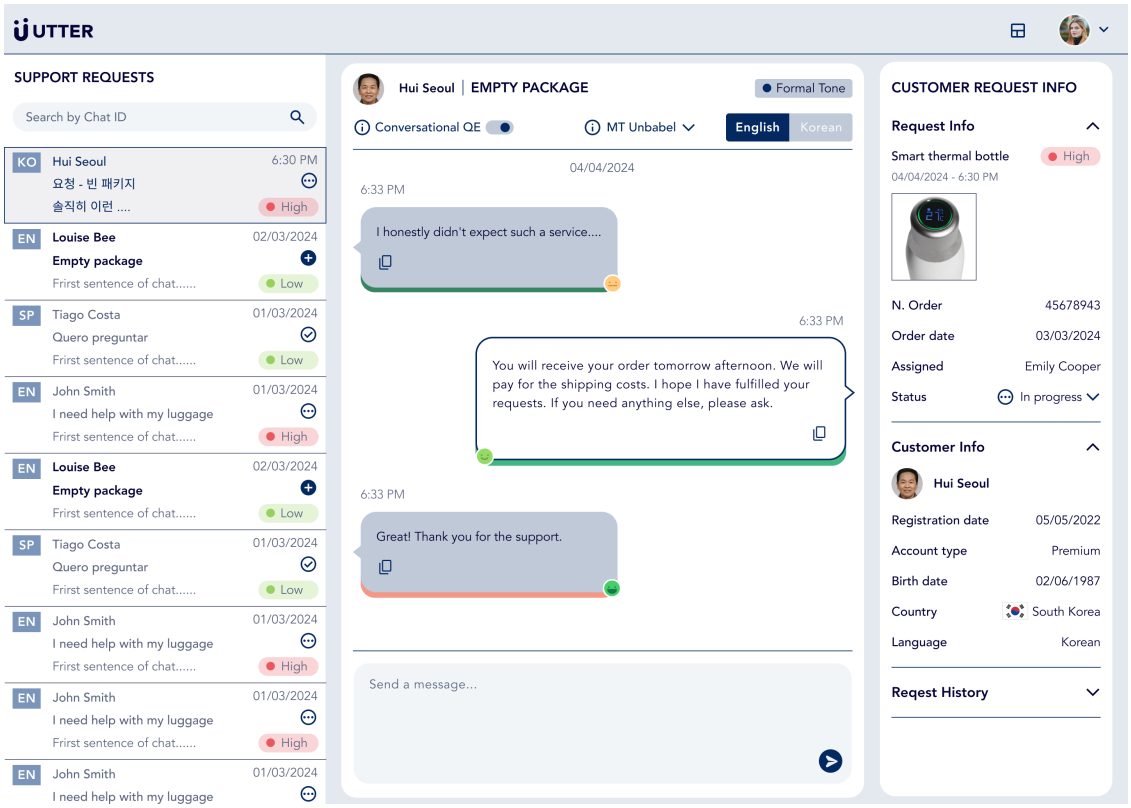


Figure 2: Customer Service Assistant – Colour code implemented.

2 Summary of Results and Plans

2.1 Results

As presented in the previous paragraphs, the final results of the PenGUIn project are the finalized prototypes of two user interfaces designed to support the UTTER use cases in terms of usability and user experience. The outputs are delivered as Figma files and demos, available at these two links:

- Customer Assistant Interface Prototype
- Meeting Assistant Interface Prototype

2.2 Business plan

In terms of business exploitation, RE:LAB will build on the experience and knowledge gathered in the context of the PenGUIn project to target opportunities in business and in research. RE:LAB will therefore seek collaborations on this topic with existing clients and partners and also by establishing new contacts. As an example of potential opportunities for collaboration resulting from the project, RE:LAB has had an introductory meeting with a research team in Naver Labs Europe to discuss potential research synergies. Leveraging its strong focus on research, RE:LAB – as part of a standard practice - will also pursue opportunities for the technical and scientific exploitation of the project results, for example by submitting scientific articles on the methodology and approach applied to this research theme.

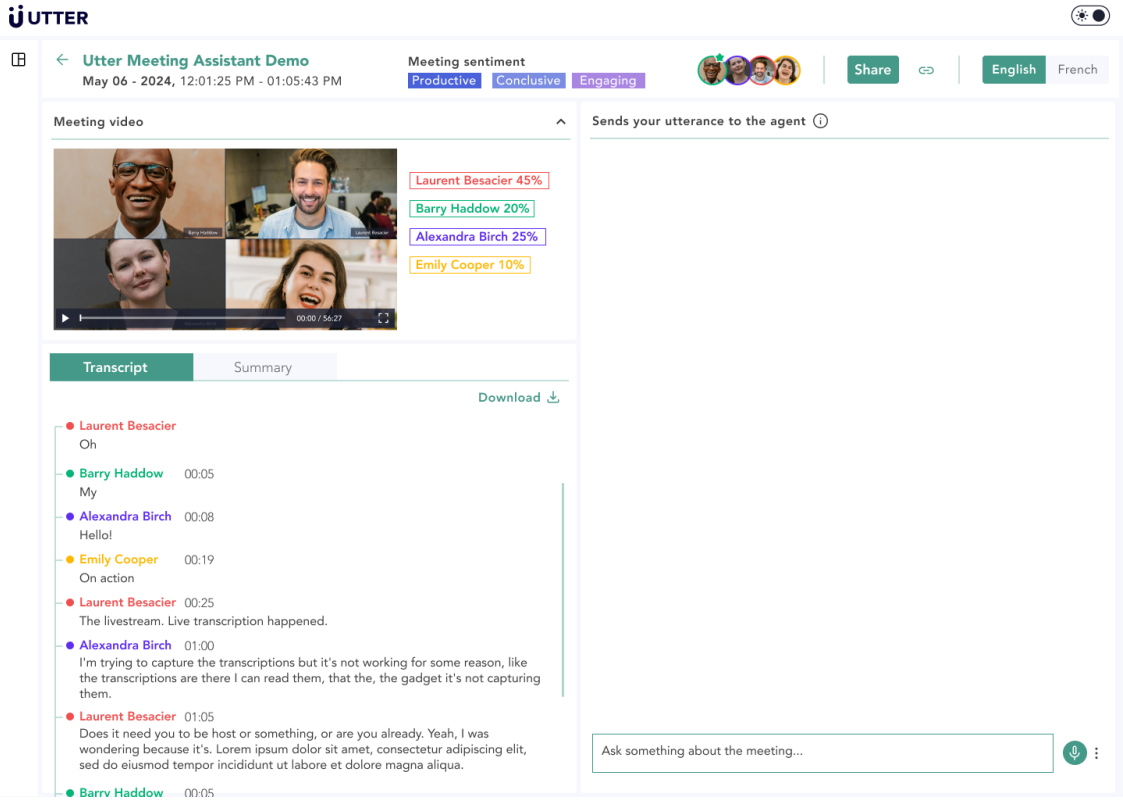


Figure 3: Use Case 2 Meeting Assistant – Light mode option

2.3 Future plans

As described in the PenGUIn project proposal, from a business perspective the project has represented an additional opportunity to strengthen our positioning in the design of innovative User Interfaces across different domains and applications, and to consolidate the expertise and RE:LAB research portfolio of technological solutions for user-centric HMIs. PenGUIn is therefore a company case study for UI and UX projects for online platforms, showcasing our goal to improve user experience in the context of automated and intelligent applications (AI assistants), and multilingual services. In terms of sustainability and exploitation, other funding avenues, from public and private sources, will be explored building on the results produced by the project, to further expand our research and development competences. In the context of Horizon Europe and other European funding programmes, RE:LAB, through its dedicated R&D team, will seek for specific calls and partnerships to leverage the PenGUIn experience and perform new research activities based on it.

2.4 Blurb for public dissemination on UTTER's website

RE:LAB was selected as one of the successful submissions to the UTTER's First Open Call with its project PenGUIn. The project, spanning 9 months, aimed to enhance user experience through an intuitive, inclusive, and adaptive Graphical User Interface (GUI) for online platforms. This was done by studying the most appropriate information design framework and applying suitable interaction strategies to support user's tasks in the context of two case studies: a customer assistant platform and an online meeting platform. PenGUIn's concept was driven by innovation and usability to achieve functionality, effectiveness, and ergonomic experience, building on RE:LAB's

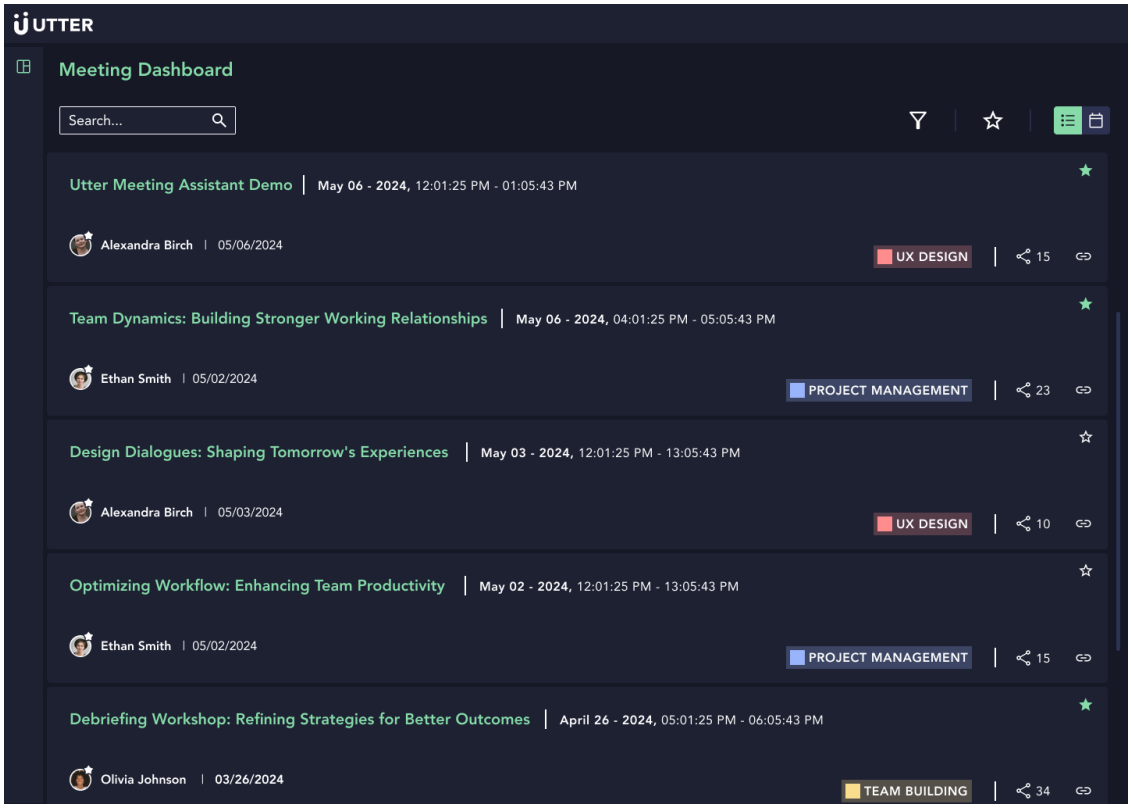


Figure 4: Meeting lines view.

user-centric methodology, "Interaction Engineering". The purpose of PenGUIn's design effort was to guide the user through the multiple platforms' functionalities, from the multilingual translation to the AI-assistant.

PenGUIn UI supported transparent and task-oriented dialogue and interaction between users of these virtual platforms. The project focused on customization flexibility, going through several design iterations, and validating the prototypes through expert analysis, focus group, and testing. The work carried out in the project has represented an additional opportunity to experiment RE:LAB original proposition and new research purposes, to consolidate the team expertise in creating and testing novel user experiences. The final prototypes are available as interactive demos at these links:

- Customer Assistant Interface Prototype
- Meeting Assistant Interface Prototype

3 Recommendation by Project Sponsor

The project proposal was aiming at "design and test a user-centred design approach to prototype a library of graphical elements that are intuitive, cross-cutting, and compatible with usability criteria" considering the UTTER use cases. The work resulted in two user interface prototypes that have been tested in focus groups to evaluate their usability. These user interfaces were made available as Figma templates that could be used as a base for developing graphical user interfaces using

any desired front-end framework. The project delivered on what has been proposed. To the best of our knowledge the project has been disseminated on social media channels and in the company's webpage. The project team documented business plans and possible future works including possible opportunities collaboration with one of the institutions that belong to UTTER (NAVER). Based on this, the recommendation is to approve the final payment to the project Awardee.

B.3 HR-XR-XTEND

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP1 Final – HR-XR-XTEND

Croatian XR Extensions

Nature	Final Report	**Work Package**	WP1
Project start date	15/01/2024	**Project end date**	15/10/2024
Interim meeting	18/06/2024	**Report submission Date**	11/10/2024
Main authors	Gaurish Thakkar, Marko Tadić		
Co-authors	Matea Filko, Daša Farkaš, Vanja Štefanec		
Reviewers	Maryam Hashemi (UVA)		
Version Control			
v0.1	**Status**	Draft	04/10/2024
v1.0	**Status**	Final	11/10/2024

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

From call documentation (section 7.2). *The final evaluation of a project will be performed by the Project Sponsor after the dissemination activities took place. The project team is required to report their results, business plans, secured venture capital for further development and future plans. The Pilot Board will assess the finished projects and evaluate the immediate results. It will also formulate recommendations for sustainability and future operation of the project. The Project Sponsor will then prepare a short report (to be made public) and recommend to the Pilot Board to approve (or not) the final payment to the project Awardee.*

How to complete this report. *The Sponsor asks the Project Team to fill in Sections 1 and 2 prior to the meeting (this entire report is likely no longer than 2–4 pages). After the meeting, the Sponsor writes a recommendation to the Pilot Board (Section 3).*

1 Project Execution

1.1 Deviations from original plan

There were no significant deviations in project execution, other than slight changes to the schedule. Data collection took more time than initially anticipated, but the effects were mitigated by employing iterative data deduplication and training methods. For the evaluation objective, we initially planned to translate the Alpaca dataset[1] into Croatian, but we dropped the idea as we found already translated Alpaca dataset into Croatian in an online repository.

Namely, in the initial steps we selected a subset from Alpaca, translated it automatically using Google Translate, and used native speakers of Croatian to proofread and adjust the translations. The texts required significant adjustments, which was impossible to complete due to the size of the dataset and available person hours. Thus we decided to use the MMLU dataset, which uses text from a wide range of different domains, and was available as already translated to Croatian using translation by GPT3. We have also compared the sample of the MMLU dataset translated with GPT3, Google Translator and the translator available at the National Language Technology Platform Hrvojka (https://hrvojka.gov.hr/). The overall accuracy of Google Translator and Hrvojka is similar. However, we opted for the GPT3 translation. Although it may be a bit worse at the individual sentence level, the overall texts, especially longer ones, appear more coherent. Lastly, it should be noted that in some cases the problem of the translation can be the result of the sample itself: the original English sentences are not actually of the highest quality, and they have a lot of noise (correct answers, incorrect answers, problems with questions etc.).

1.2 Development

Objective 1. Collecting the training corpus

For the purpose of training the monolingual Croatian LLM, a large-scale data set was composed from the available monolingual corpora of Croatian language, parallel corpora containing Croatian as one of the languages, as well as several multilingual corpora composed of, other than Croatian, closely related South-Slavic languages, i.e., Serbian, Bosnian, and Montenegrin. Texts in other languages were filtered out of the multilingual corpora, and only texts in Croatian were used. Filtering was performed using the available metadata assigned to corpus samples. In some cases,

[1] https://huggingface.co/datasets/yahma/alpaca-cleaned

samples were already labelled for language, and in others, other metadata attributes were used, such as the domain URL. In Table 1, we list down all the data sources.

Name	Approx Size
CLASSLA Hr Web corpus 1.0	2.5 billion
CC100-Hr Dataset	2.27 billion
Corpus of Croatian News Feeds	2.25 billion
Parallel data for En-Hr on OPUS Resources*,	1.48 billion
Hr-news from XLM-R-BERTić dataset	1.4 billion
Croatian news/legal corpus	175 million
Corpus of Croatian Academic Theses	312 million
ParaCrawl*	69.96 million
Riznica from XLM-R-BERTić dataset	69.51 million
MARCELL Croatian legislative subcorpus	56 million
CURLICAT Croatian corpus	49 million
MARCELL Croatian-English Parallel Corpus of Legislative Texts*	14.3 million
Romance-Croatian Parallel Corpus* (literary works)	2.5 million
Total	8.9 billion

Table 1: Non-exhaustive list of largest data sources used for training the HR-GPT (Beta version) with approximate size in tokens. *Croatian texts only

We used the datatrove (Penedo et al., 2024) library to perform the near deduplication with MinHashLSH and a threshold of 0.72, following the advice that LLMs trained on deduplicated data are better and memorise less of their data (Lee et al., 2022). After deduplication, the deduplicated dataset is approximately 7.72B tokens in size, compared to the original dataset, which contains 8.9B tokens. The dataset was divided into a training and evaluation and test subset in 94:5:1 ratio.

Objective 2. Training the language model

We divided the language model training into three cases.

1. **Training from scratch:** In this case, we trained the model using the existing training configurations from the "GPT-NeoX" library (Andonian et al., 2023). We chose the following parameters based on the number of tokens available for training: 160M, 350M, 410M, and 1.4B parameters. The models are based on the GPT-2 tokenizer. We trained an additional model with 160M parameters, which relied on a newly trained tokenizer (hr-tok) from the training set. We conducted this study to examine the impact of a tokenizer specifically trained on Croatian texts.

2. **Continued pretraining on the monolingual model:** We investigated the effect of continued pretraining on the publicly available GPT-2[2] model. The model is originally monolingual, and we trained using the same Croatian text as in the previous scenario.

[2] https://huggingface.co/openai-community/gpt2

3. **Continued pre-training on multilingual model:** To perform continued pretraining, we used the quantised version of Gemma 7b, i.e., "unsloth/gemma-7b-bnb-4bit". We investigated the results of using the existing multilingual large language model as a backbone for further training.

Objective 3. Evaluation

There are three parts to the evaluation:

- Benchmark datasets for zero-shot evaluation
- Supervised instruction tuning with the Alpaca dataset
- Sentiment and choice of plausible alternatives datasets for supervised fine-tuning.

We conducted the evaluation using the evaluation-harness library (Gao et al., 2024). We used the following benchmarks: TruthfulQA (Clark et al., 2018), Multilingual ARC (Clark et al., 2018), Belebele (Bandarkar et al., 2024), Multilingual HellaSwag (Zellers et al., 2019), and the MMLU (Hendrycks et al., 2021a,b) dataset. In addition, we used two tasks from the Benchich[3] benchmarking dataset, namely sentiment analysis (SA) and choice of plausible alternatives (COPA). The sentiment task associated with sentiment identification in parliamentary proceedings and COPA evaluates cause and effect of premise and hypothesis in Croatian. We conducted evaluations in a zero-shot, three-shot, and ten-shot setting. Two linguists manually checked the MMLU dataset (en) and its Croatian translations, which the University of Oregon had translated using GPT-3.5-turbo[4]. We compared a total of 150 samples (75 per linguist) by checking them with their corresponding Google translations. Additionally, we performed supervised fine tuning on the Alpaca dataset (Croatian version) and evaluated the trained models using the benchmarking datasets.

No supervised training (zero-shot evaluation)										
		Pretraining					Vanilla		CPT	
benchmark	metric	160M	350M	410M	1.4B	160M+hrtok	gpt2	gemma-7b	gpt2-en-cpt-hr	gemma-7b-cpt
arc_hr	acc	18.91	20.96	20.36	20.44	20.87	19.85	**32.34**	18.82	21.81
	acc_norm	23.44	25.49	25.06	24.89	23.95	23.87	**36.53**	23.44	24.55
belebele_hrv_Latn	acc	22.78	23	23.11	22.67	22.78	23.44	**52.67**	21.33	23
	acc_norm	22.78	23	23.11	22.67	22.78	23.44	**52.67**	21.33	23
hellaswag_hr	acc	28.43	29.87	30.08	31.36	28.63	26.27	**38.5**	26.44	24.38
	acc_norm	30.07	32.74	33.38	35.52	30.63	29.42	**50.11**	28.14	24.24
m_mmlu_hr	acc	22.65	25.21	22.8	22.54	22.63	22.59	**41.5**	22.67	25.02
truthfulqa_hr_mc1	acc	25.88	24.58	25.75	26.27	25.49	22.24	**28.61**	26.01	18.34
truthfulqa_hr_mc2	acc	43.82	42.21	42.34	42.52	43.03	40.8	46.6	**46.79**	-

Table 2: Benchmarking evaluation (zero-shot) results for a variety of models without the use of any supervised training. The table displays scores for various models that did not utilise any supervised training (instruction fine tuning). ACC: accuracy and acc_norm: normalised accuracy.

The following are the key observations:

[3] https://github.com/clarinsi/benchich
[4] https://huggingface.co/datasets/alexandrainst/m_mmlu

Trained on benchich training data										
dataset	metric	160M	350M	410M	1.4B	160M+hrtok	gpt2	gemma-7b	gpt2-en-cpt-hr	gemma-7b-cpt
SA-Parlasent(hr-only)	acc	68.86	72.98	72.53	71.03	71.18	36.68	72.46	53.74	**74.48**
COPA	acc	50	49.4	49.6	47.8	48.8	48.4	**79.8**	50.2	79.6

Table 3: The model scores (accuracy) for supervised tasks related to sentiment analysis and choice of plausible alternatives (COPA).

Supervised training (instruction fine tuning)									
Alpaca									
benchmark	160M	350M	410M	1.4B	160M+hrtok	gpt2-en-cpt-hr	gemma-7b-cpt	gpt2	gemma-7b
arc_hr	21.21	19.76	22.75	23.1	20.19	19.08	35.76	19.67	**35.93**
	26.26	25.32	25.75	27.12	23.18	24.64	37.81	24.12	**39.95**
belebele_hrv_Latn	22.67	22.89	23.44	22.67	22.67	23.78	**58**	23.89	45.33
	22.67	22.89	23.44	22.67	22.67	23.78	**58**	23.89	45.33
hellaswag_hr	28.56	30.14	30.74	31.64	28.96	26.84	40.35	26.27	**41.1**
	30.19	32.76	33.75	35.46	30.39	27.62	53.56	27.7	**53.67**
m_mmlu_hr	22.69	23.05	22.82	22.76	22.79	22.63	**43.12**	22.62	33.12
truthfulqa_hr_mc1	24.19	22.63	23.67	26.92	25.23	25.1	**31.73**	24.45	30.04
truthfulqa_hr_mc2	42.58	40.8	39.2	42.87	42.93	41.1	**50.04**	40.08	47.68

Table 4: Benchmarking evaluation results for a variety of models trained with the Alpaca instruction tuning dataset.

1. **Performance Variation Across Models:** Larger models like "gemma-7b" tend to perform better than smaller ones in most tasks, both before and after Alpaca fine-tuning. For example, gemma-7b-cpt achieves high accuracy on belebele_hr, m_mmlu_hr and truthfulqa_hr tasks, indicating better performance with more complex pretraining strategies.

2. **Impact of Fine-Tuning:** Alpaca fine-tuning (+ALP) often improves performance, as seen with gemma-7b-cpt+ALP, which achieves the highest accuracy on several tasks.

3. **Comparison Between Vanilla and CPT:** Models with CPT generally show enhanced performance over their Vanilla counterparts, suggesting that CPT might be more effective for these tasks.

1.3 Dissemination

Dissemination activities were organised according to the project plan. First the project logo and website design was produced.

Figure 1: HR-XR-XTEND project logo

The project webpage was opened under the FFZG domain: https://hr-xr-xtend.ffzg.unizg.hr. The relevant project news were published on the webpage, particularly the news about the project presentation at different conferences.

The project and its results were presented at the following conferences:

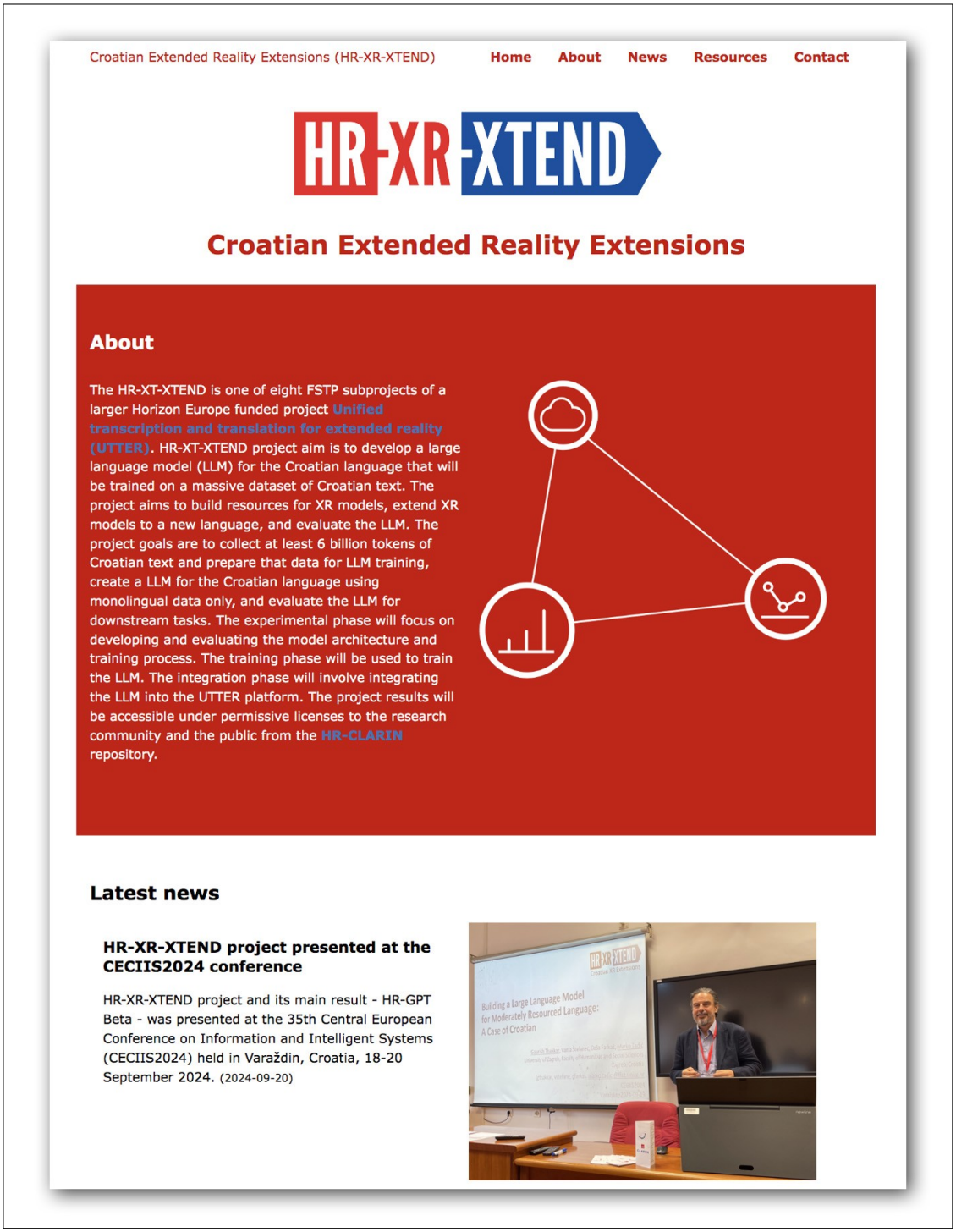


Figure 2: Snapshot of the project webpage

- **Dani e-infrastrukture 2024 (DEI) / Days of e-infrastructure 2024** (https://dei.srce.hr), Zagreb, Croatia, 16th, 18th and 19th April 2024, organiser: University of Zagreb Computing Centre. The project was presented by a poster [https://dei.srce.hr/2024/izabrane-poster-prezentacije];

- **Joint conference on Language Resources and Evaluation and Computational Linguistics (LREC-COLING2024)**, [https://lrec-coling-2024.org/], Turin, Italy, 20th to 25th May

2024, organiser: European Language Resources Association. The project was presented with a paper and poster presentation of an experiment in Sentiment Analysis [https://aclanthology.org/2024.lrec-main.946/];

- **New Trends in Translation Technology (NeTTT2024)**, [https://nettt-conference.com/], Varna, Bulgaria, 3th to 65h July 2024, organisers Lancaster University, UK and Association for Computational Linguistics, Bulgaria. The project was presented with a paper and oral presentation [https://acl-bg.org/proceedings/2024/NeTTT%202024/pdf/2024.nettt-1.17.pdf].
- **Central European Conference on Inteligent Information Systems (CECIIS2024)**, [https://ceciis.foi.hr/], Varaždin, Croatia, 18th to 20th September 2024, organiser: University of Zagreb, Faculty of Organisation and Informatics in Varaždin. The project was presented with a paper and oral presentation.
- **Festival of Languages 2024**, [https://croatia.representation.ec.europa.eu/events/festival-jezika-2024-2024-09-27_hr], Zagreb, Croatia, 27th September 2024, organiser: Representation of the European Commission in Croatia. The project was presented with an oral presentation.
- **Metaphorical Collocations (MetaKOl2024)**, [https://metakol.uniri.hr/en/konferencija-2024/], Dubrovnik, Croatia, 3rd and 4th October 2024, organiser: University of Rijeka. The HR-GPT Beta was presented with an oral presentation.
- **21st EURALEX International Congress Lexicography and Semantics (EURALEX2024)**, [https://euralex.jezik.hr/], **Workshop Large Language Models and Lexicography (LLM-Lex)**, Cavtat, Croatia, 8th to 12th October 2024, organiser: Institute for the Croatian Language. The HR-GPT Beta was presented by the oral presentation [https://www.cjvt.si/en/research/community/llm-lex-2024/].

1.4 Ethics

Since the training data set was composed from publicly available data, we expect that ethical issues have been sorted out by the data providers. All results of the project available under permissive licences in the HR-CLARIN repository, will always have a link to the original source of data and possible users will be able to check the status of ethical issues directly at the data source.

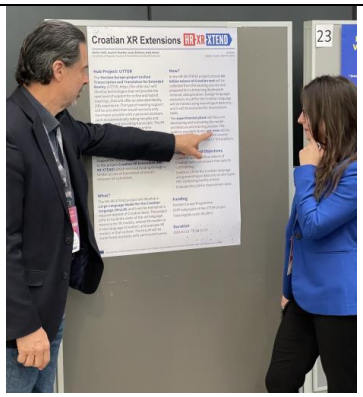
2 Summary of Results and Plans

2.1 Results

The project results and training data will be available at the Croatian CLARIN repository (https://www.clarin.hr/) under permissive licenses by the end of the project.

The key results of the project are:

- **HR-GPT Beta** trained in four sizes: 160M, 350M, 410M i 1.4B parameters;
- **cleaned training data set for HR-GPT Beta**, available only partially because for some of the training data we couldn't reach an agreement with data providers about distribution to the third parties;



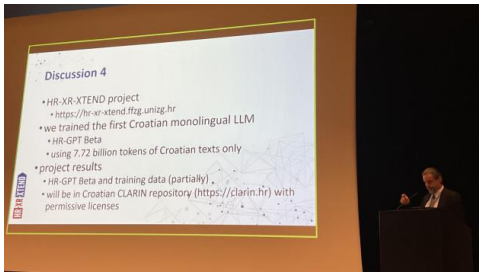
DEI 2024

LREC-COLING 2024

NeTTT 2024

CECIIS 2024

MetaCol 2024



EURALEX 2024

Figure 3: Photos of the project presentations at conferences

- **four sets of training data** for training the high-precision language identifier between Bosnian, Croatian, Montenegrin and Serbian languages (100 million tokens for each language). This language identifier will be trained by the Charles University, a coordinator of the High Performance Language Technology (HPLT) project.

2.2 Business plan

The fundamental prerequisite for the usage of HR-GPT Beta in different business environments is its availability in an digital repository with persistent identifiers. This prerequisite is fulfilled by depositing project results in the HR-CLARIN repository. The permissive licenses will allow open access usage of the results by researchers and developers for different purposes.

Since the research team has good connections with a number of similar research teams in Europe, we are aware of interest for our results that has been expressed already by different projects, e.g. Charles University with HPLT project, Tilde within the Large AI Grand Challenge, etc.

Our previous collaboration with a number of translation and localisation companies in Croatia opens also the possibility of deployment of the HR-GPT Beta in the post-processing of the MT output. The company Ciklopea Ltd. already expressed their interest in inclusion of this LLM in their work process.

2.3 Future plans

The future work can be divided into two directions.

In the first direction, we will collect and clean additional data for Croatian and perform additional filtering in terms of quality, but with much slower pace since the funding from HR-XR-XTEND expires. The non-exhaustive list of already collected and processed data combined with available additional data is presented in the Table 5.

In the second part, we would like to evaluate the models for various other NLP tasks like autocompletion and error correction.

Also, since one of the models we trained didn't converge on train loss for the 160M parameter model's default training configuration, more research is necessary for models that use the "hr-tok" tokenizer. We believe additional training is required.

2.4 Blurb for public dissemination on UTTER's website

The "Croatian XR Extensions" project aimed to create a large-scale monolingual Croatian language model (HR-GPT Beta). A significant training dataset was collected and cleaned from existing mono- and multilingual resources that include texts in Croatian. The preprocessing featured also advanced deduplication techniques, resulting in a final training dataset of 7.72 billion tokens. Three training scenarios were used: training from scratch, continued pretraining on a monolingual model, and continued pretraining on a multilingual model. The evaluation was performed using several benchmark datasets, and fine-tuning with the Alpaca dataset improved model performance. Larger models, like "gemma-7b", outperformed smaller ones, and fine-tuning enhanced results further. Key results include multiple model versions (160M, 350M, 410M, and 1.4B parameters) and a cleaned training dataset. Future work involves additional data collection, additional

Name	Approx Size
CC100-Croatian Dataset 1.0	3.3 billion
CLASSLA Hr Web corpus 1.0	2.5 billion
Corpus of Croatian News Feeds	2.25 billion
HPLT Croatian/Bosnian/Serbian Corpus, Croatian texts only	4 billion
Parallel data for En-Hr on OPUS Resources, Croatian texts only	1.48 billion
Corpus of Croatian Academic Theses	312 million
Joel Niklaus, Multi Legal Pile, Croatian	258 million
Leipzig Corpora	182.40 million
ParlaMint 4.0, Croatian texts only	88.16 million
ParaCrawl, Croatian texts only	79.06 million
hrWikipedia	66.48 million
MARCELL Croatian legislative subcorpus	56 million
Total	14.57 billion

Table 5: Non-exhaustive list of already collected data sources with approximate size in tokens for sources with 50+ million tokens.

model training, further NLP task evaluations, and more training experiments. The HR-GPT Beta and training material (partially) will be publicly accessible under permissive licenses from the HR-CLARIN repository (https://clarin.hr). More information can be found on the project website https://hr-xr-xtend.ffzg.unizg.hr.

3 Recommendation by Project Sponsor

This space is intended for the assessment by the project Sponsor. Did the project deliver its planned results? Was it successfully disseminated? Did the project team document business plans and future plans, and do these seem appropriate? Are there opportunities for exploitation of results within UTTER?

The project planned the collection of Croatian datasets and training a large language model on Croatian and they delivered on this objective. This project was successfully disseminated. UTTER could use these datasets for training the EuroLLM langaugae model. This project has completed successfully.

References

Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Jason Phang, Shivanshu Purohit, Hailey Schoelkopf, Dashiell Stander, Tri Songz, Curt Tigges, Benjamin Thérien, Phil Wang, and Samuel Weinbach. GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch, 9 2023. URL https://www.github.com/eleutherai/gpt-neox.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.44.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024. URL https://zenodo.org/records/12608602.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021b.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577. URL https://aclanthology.org/2022.acl-long.577.

Guilherme Penedo, Hynek Kydlíček, Alessandro Cappelli, Mario Sasko, and Thomas Wolf. Datatrove: large scale data processing, 2024. URL https://github.com/huggingface/datatrove.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

B.4 SignReality

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP1 Final – SignReality

Extended Reality for Sign Language translation

Nature	Final Report	**Work Package**	WP1
Project start date	dd/mm/2024	**Project end date**	dd/mm/2024
Interim meeting	dd/mm/2024	**Report submission Date**	dd/mm/2024
Main authors	Sponsor (PARTNER)		
Co-authors	Awardees (ORG)		
Reviewers	Maryam Hashemi (UVA)		
Version Control			
v0.1	**Status**	Draft	XXX
v1.0	**Status**	Final	XXX

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Project Execution

1.1 Deviations from original plan

The following deviations have occurred at the development:

- The API for the communication between the augmented reality devices, the animation engine and the translation engine (WP2) is not fully implemented due to platform-specific incompatibilities. Back-off solution: the users have to click on Hololens 2 to play pre-recorded animations, generated at the animation engine at an earlier stage.
- The animation loading time on Hololens 2 (WP1) is unsuitable for real-time communication.
- XReal implementation (WP1) remains prototypical, and it is not possible to invoke animations, due to lack of documentation and the steep learning curve.
- Open source licensing and public distribution is not available for all parts of the pipeline and corpora due to licensing reasons beyond the control of the project and could not be solved within such a short duration. We are working to resolve them in near future.
- Scientific publications will be submitted in the near future due to the short project duration

Key results are ready (translation module, animation framework, avatar adjustment, evaluation study, demo). We see the above issues rather as software engineering problems than a scientific problem.

1.2 Development

1.2.1 WP1: Avatar animation and representation

Avatar animation engine During the development of the SignReality prototype, we have developed two new features of our sign language synthesis system[1]:

- We have developed a remote HTTP API for remote submission of MMS data and retrieval of avatar animation data. Essentially, the main rendering engine has been wrapped in a Flask server that receives MMS instances and runs the animation engine. The result, a JSON file containing a full animation of the SL sentence, is then returned.
- We improved the quality of the motion synthesis by revising the coordinate systems of the inflection of the hand motion. Originally, the motion of the hands was inflected relatively to the avatar center of the body. After several experiments and observation, we realized that it was going to be more intuitive and stable to perform inflections relatively to its torso.

[1] A demo can be found at:

Avatar representation Our work involves the display of the avatar on two XR devices:

- **Hololens 2**: We have improved prior implementation by adding user interaction features, through which the user can move and place the avatar in the augmented space. Such positioning mechanism has then been used to run the user studies.
- **XReal Light**: We ported a pre-existing Unity-mobile version to XReal Light, due to its low weight and cost. The implementation (incl. Android app) has remained into a prototypical stage. The avatar has been ported, but the animations have not been connected to a control panel, and therefore it is not possible to invoke them.

1.2.2 WP2: Translation model from text to sign language representation

Implementation of translation model An encoder-decoder model has been trained on the AVASAG corpus in the domain of train announcements (Nunnari et al., 2021; Bernhard et al., 2022), by continual learning of an NLLB model[2] (Team et al., 2022). In order to support MMS annotation as a supplementary annotation over the produced glosses, we have developed a custom PyTorch-based code, which also allows for executing an XML-RPC server to provide translations on demand. Although during the development, the model exhibited high performance (>30 BLEU score) on a limited test set of similar domain with cross-validation, small-scale human evaluation indicated that the model hallucinates occasionally and may have overfitted due to the small training corpus.
Results: basic model, XML-RPC server

Corpus acquisition and curation Due to the lack of existing resources, we continued fundamental research on making more parallel corpora available and usable. We have collected big amounts of subtitled sign language TV shows (DGS, various South American sign languages) from online sources and are acquiring permissions to use them for research, and we have also been recording our own corpus of German fairy tales, as a continuation of recent effort (DGS-Fabeln-1 Nunnari et al., 2024). We worked on two methods for aligning video (signed content) with subtitles:

1. Sign language segmentation based on temporal features (optical flow; Kishore et al., 2016; Zhang et al., 2019). The first ever sentence-level segmentation models were built for BSL and ASL.
   Result: The work is being finalized and will be submitted to an academic venue.
2. Video-text alignment based on i3D features using a Transformer (Bull, 2023; Bull et al., 2021), with the aim to fine-tune existing models for BSL to DGS and other sign languages.
   Result: data preprocessing, human annotation, first statistics on the average timeshift.

1.2.3 WP3: Participatory design and evaluation

Extending previous research (Nguyen et al., 2021; Nolte et al., 2022, 2023), we performed a usability study with the participation of 9 fluent speakers of the German Sign Language (8 of them deaf)

[2] Pre-trained NLLB model: `nllb-200-distilled-600M`

at the Centre for Culture and Visual Communication of the Deaf Berlin Brandenburg (ZFK). The study lasted about 9 hours and focused on *intelligibility, user experience* and *acceptance* focusing on HoloLens 2, and particularly the user-based interaction mechanisms to adjust and position the avatar in space. The participants tried out the avatar application on the HoloLens 2 for themselves in two sessions with variable scenaria. Every session was followed by a questionnaire following the *Raw Task Load Index* (TLX; Hart and Staveland, 1988) and the *Short Version of the User Experience Questionnaire* (UEQ-S; Schrepp et al., 2017). Finally, we collected the participants' opinions in the form of a short interview. According to the study, the currently implemented positioning features did not provide better user experience, but the users insisted on the need of the adjustment features and indicated valuable feedback for further improvements, such as difficulty with the gestures, unpractical size of the avatar after placement, missing interaction feedback from the control panel.
Result: Analysis, publication of results in a CHI-related conference.

1.3 Dissemination

The following dissemination tasks have been completed:

- **Community engagement:** Communication of the idea and user study at ZFK. Internal newsletter post at DFKI. Presentation of project concept and participation at the network of the Berlin XR Lab[3].
- **Web pages:** Project descriptions at the DFKI website[4], department websites[5] and personal websites[6].
- **Social media**: News about the project and user evaluation posted on the Affective Computing group LinkedIn page[7].
- **Academic activities:** The project is aligned (partially or entirely) with one BSc thesis, 4 MSc theses and one student coding workshop at the Technical University of Berlin and the Saarland University. Upcoming publications for text segmentation, sign language animation, evaluation study.

1.4 Ethics

Our present experiments on DGS are part of a broader research aiming to provide equal access to language technology for sign language users. The users of a sign language form a linguistic and cultural minority and the fact that the project is led by researchers that are hearing people entails the risk of developments that are not in accordance with the will of the former and lead to complaints for cultural misuse and appropriation. For this reason, in our broader research we have included interpreters and members of the deaf and hard of hearing communities as part of the

[3] Berlin XR Lab: https://www.berlin-xrlab.de/
[4] DFKI website: https://www.dfki.de/en/web/research/projects-and-publications/project/signreality
[5] Design Research Lab: https://www.dfki.de/en/web/research/projects-and-publications/project/signreality, Berlin Open Lab: https://berlin-open-lab.org/portfolio/signreality-extended-reality-for-sign-language-translation/
[6] Fabrizio Nunnari: https://www.dfki.de/~fanu01/
[7] LinkedIn: https://www.linkedin.com/company/affective-computing-group2021/

research team, consultants and participants in user studies and workshops, and we have been in constant co-operation with related unions and communication centers.

DGS is analysed and depicted is in a preliminary stage and results should by no means be presented as a functional product, which the respective communities might find offensive and diminishing. In particular, glosses are known to be inferior to the full linguistic capacity of the sign languages and are only seen as a methodological tool to aid further research. The signing avatar is lacking several elements of the sign language (smooth hand movements, facial expressions, mouthings) and should be seen as work in progress.

The users have been informed and consented as per GDPR about the storage of personal information and the video-recordings of the user study (which will only be processed internally to the project and won't be published). The results are anonymized. Users have participated in the study as part of their working time in ZFK and additionally received a compensation coupon. ZFK was compensated with a lump sum. A DGS interpreter was available for the entire duration of the study.

2 Summary of Results and Plans

2.1 Results

The results of the project have been uploaded to a cloud folder.[8] They include:

- Translation server and trained model (code and model)[9]
- Full translation pipeline with avatar animation engine (code and demo)
- Hololens 2 implementation with adjustment features (code and demo)
- XLReal Light port of the avatar, Android app (code)
- Sign language segmentation method (report)
- User study results and feedback (report)

2.2 Future plans

The results of the research will be reformatted as academic papers and will be submitted to relevant venues (e.g. *ACL, CHI, IVA, SLTAT). Successful components will be extended and integrated into relevant research projects (e.g. BIGEKO, Federal German Ministry of Education and Research, 2023-2026). Further research funding will be sought from European, federal and industrial sources.

2.3 Blurb for public dissemination on UTTER's website

The project "SignReality" achieved significant milestones in bridging sign language technology with Extended Reality. Key results include the development of an engine for avatar animation, accompanied by device-specific implementation on two AR devices (Hololens 2 and XReal Light). Translation from spoken language to a textual sign language representation (German→DGS) was enabled through an encoder-decoder translation model, whereas further improvement of relevant

[8] Project results: https://cloud-affective.dfki.de/s/o5SCGE8JZn4wsfN
[9] Translation server code: https://github.com/DFKI-SignLanguage/text-to-gloss-machine-translation

models will benefit from the work on corpus acquisition and alignment. The implementation was tested for *intelligibility, user experience* and *acceptance* in a user study with native sign language users at the Centre for Culture and Visual Communication of the Deaf Berlin Brandenburg (ZFK) providing valuable feedback. The project has been integrated into several academic theses and university workshops, and research findings will be submitted in relevant academic venues.

3 Recommendation by Project Sponsor

The project clearly achieved all of its goals, with very minor deviations from the original plan, both along the scientific and dissemination dimensions. The project team has experience with the ethical considerations behind the experimental setup and did a remarkable job both at complying with all relevant guidelines and regulations but also at clearly documenting the scope of their findings and technology. I recommend payment of the second installment.

Glossary

ASL American Sign Language. 4

BSL British Sign Language. 4

DGS German Sign Language (Deutsche Gebärdensprache). 4–6

MMS Multimodal Signstream, (consists of the annotated sentences augmented with the sign inflection parameters). 3, 4

References

Lucas Bernhard, Fabrizio Nunnari, Amelie Unger, Judith Bauerdiek, Christian Dold, Marcel Hauck, Alexander Stricker, Tobias Baur, Alexander Heimerl, Elisabeth André, Melissa Reinecker, Cristina España-Bonet, Yasser Hamidullah, Stephan Busemann, Patrick Gebhard, Corinna Jäger, Sonja Wecker, Yvonne Kossel, Henrik Müller, Kristoffer Waldow, Arnulph Fuhrmann, Martin Misiak, and Dieter Wallach. Towards Automated Sign Language Production: A Pipeline for Creating Inclusive Virtual Humans. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '22, pages 260–268, New York, NY, USA, July 2022. Association for Computing Machinery. ISBN 978-1-4503-9631-8. doi: 10.1145/3529190.3529202. URL https://doi.org/10.1145/3529190.3529202.

Hannah Bull. *Learning sign language from subtitles*. PhD thesis, Université Paris-Saclay, Paris, France, 2023. URL https://theses.hal.science/tel-04055873v1/file/112750_BULL_2023_archivage.pdf.

Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni, and Andrew Zisserman. Aligning Subtitles in Sign Language Videos. *Proceedings of the IEEE International Conference on Computer Vision*, pages 11532–11541, May 2021. ISSN 15505499. doi: 10.1109/ICCV48922.2021.01135. URL https://arxiv.org/abs/2105.02877v1. arXiv: 2105.02877 Publisher: Institute of Electrical and Electronics Engineers Inc. ISBN: 9781665428125.

Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988. URL https://www.sciencedirect.com/science/article/pii/S0166411508623869.

P.V.V. Kishore, M.V.D. Prasad, D. Anil Kumar, and A.S.C.S. Sastry. Optical Flow Hand Tracking and Active Contour Hand Shape Features for Continuous Sign Language Recognition with Artificial Neural Networks. In *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, pages 346–351, February 2016. doi: 10.1109/IACC.2016.71. URL https://ieeexplore.ieee.org/abstract/document/7544860.

Lan Thao Nguyen, Florian Schicktanz, Aeneas Stankowski, and Eleftherios Avramidis. Evaluating the translation of speech to virtually-performed sign language on AR glasses. In *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 141–144, June 2021. doi: 10.1109/QoMEX51781.2021.9465430. URL https://ieeexplore.ieee.org/abstract/document/9465430. ISSN: 2472-7814.

Amelie Nolte, Karolin Lueneburg, Dieter P. Wallach, and Nicole Jochems. Creating Personas for Signing User Populations: An Ability-Based Approach to User Modelling in HCI. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '22, pages 1–6, New York, NY, USA, October 2022. Association for Computing Machinery. ISBN 978-1-4503-9258-7. doi: 10.1145/3517428.3550364. URL https://doi.org/10.1145/3517428.3550364.

Amelie Nolte, Barbara Gleißl, Jule Heckmann, Dieter Wallach, and Nicole Jochems. "I Want To Be Able To Change The Speed And Size Of The Avatar": Assessing User Requirements For Animated Sign Language Translation Interfaces. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, pages 1–7, New York,

NY, USA, April 2023. Association for Computing Machinery. ISBN 978-1-4503-9422-2. doi: 10.1145/3544549.3585675. URL https://doi.org/10.1145/3544549.3585675.

Fabrizio Nunnari, Judith Bauerdiek, Lucas Bernhard, Cristina España-Bonet, Corinna Jäger, Amelie Unger, Kristoffer Waldow, Sonja Wecker, Elisabeth André, Stephan Busemann, Christian Dold, Arnulph Fuhrmann, Patrick Gebhard, Yasser Hamidullah, Marcel Hauck, Yvonne Kossel, Martin Misiak, Dieter Wallach, and Alexander Stricker. AVASAG: A German Sign Language Translation System for Public Services (short paper). In Dimitar Shterionov, editor, *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 43–48, Virtual, August 2021. Association for Machine Translation in the Americas. URL https://aclanthology.org/2021.mtsummit-at4ssl.5.

Fabrizio Nunnari, Eleftherios Avramidis, Cristina España-Bonet, Marco González, Anna Hennes, and Patrick Gebhard. DGS-fabeln-1: A multi-angle parallel corpus of fairy tales between German Sign Language and German text. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4847–4857, Torino, Italy, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.434.

Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). 2017. ISSN 1989-1660. doi: 10.9781/ijimai.2017.09.001. URL https://idus.us.es/handle/11441/107084. Accepted: 2021-04-14T11:12:39Z Publisher: UNIR.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No Language Left Behind: Scaling Human-Centered Machine Translation, August 2022. URL http://arxiv.org/abs/2207.04672. arXiv:2207.04672 [cs].

Shujun Zhang, Weijia Meng, Hui Li, and Xuehong Cui. Multimodal Spatiotemporal Networks for Sign Language Recognition. *IEEE Access*, 7:180270–180280, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2959206. URL https://ieeexplore.ieee.org/abstract/document/8932517. Conference Name: IEEE Access.

B.5 DeMINT

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP1 Final – DeMINT

Automated Language Debriefing for English Learners via AI Chatbot Analysis of Meeting Transcripts

Nature	Final Report	**Work Package**	WP1
Project start date	15/01/2024	**Project end date**	15/10/2024
Interim meeting	31/05/2024	**Report submission Date**	27/09/2024
Main authors	Juan Antonio Pérez-Ortiz (University of Alicante)		
Co-authors			
Reviewers			
Version Control			
v0.1	**Status**	Draft	06/09/2024
v1.0	**Status**	Final	27/09/2024

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Project Execution

This is the final report of project DeMINT ("Automated Language Debriefing for English Learners via AI Chatbot Analysis of Meeting Transcripts"). The project is funded after a Financial Support for Third Parties (FSTP) call[1] by the EU project UTTER ("Unified Transcription and Translation for Extended Reality").[2] DeMINT started on January 15, 2024 and is expected to end on October 15, 2024.

1.1 Deviations from original plan

There are no substantial deviations from the original plan. As a minor deviation, we can mention a low-level technical issue such as the decision to not use LanguageTool for grammar checking as the open source version was not as accurate enough for our purposes. This was not a big issue, as there exist tools based on neural models to replace it.

1.2 Development

DeMINT has developed a prototype of a conversational system designed to enhance non-native English speakers' language skills through post-meeting analysis of the transcriptions of video conferences in which they have participated. Following recent advances in chatbots and agents based on large language models (LLMs), the tutoring system leverages pre-trained LLMs within an ecosystem that integrates different techniques, including in-context learning, external non-parametric memory retrieval, efficient parameter fine-tuning, grammatical error correction models, and error-preserving speech synthesis.

In addition to the research team of faculty members, 3 graduate computer science students have joined the project as technicians.

A pilot human evaluation was designed and carried out under the supervision (and after the approval) of the University's ethics committee.[3] For the pilot study, 7 students from different degrees in our University and with a level of English proficiency of B2 or C1 were selected. The students were paid for their participation according to the budgeted amount. Each student filled a previous questionnaire to collect sociodemographic information and data regarding their familiarity with information and communication technologies and chatbots.

Each student was asked to engage in 5 videoconferences with other students. In total, 10 videoconferences with two participants and 5 videoconferences with three participants were held. To provide a topic for videoconferences, we proposed a role play extracted from a popular manual for English as a second language for each of them.[4] Finally, students were asked to interact with the chatbot to have a debriefing corresponding to each of the videoconferences in which they partitipated; participants finally provided feedback on their user experience by means of a questionnaire which can be found in appendix B.

[1] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/competitive-calls-cs/3722
[2] Grant agreement number 101070631.
[3] https://web.ua.es/en/vr-investigacio/comite-etica
[4] Pitts, L. (2015). ESL Role Plays: 50 engaging role plays for ESL and EFL classes. ECQ Publishing.

1.3 Dissemination

Three main dissemination activities have been carried out so far.

- First, the main components of the system were presented at UTTER 2nd User Day, an online event held on July 5, 2024. The video of DeMINT's presentation is available on YouTube.[5]
- A paper describing the chatbot with the title "A Conversational Intelligent Tutoring System for Improving English Proficiency of Non-Native Speakers via Debriefing of Online Meeting Transcriptions" has been accepted to the 13th workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL). The paper provides a detailed description of the system's architecture. It will be presented as an oral presentation during the conference to be held in Rennes (France) on 25–26 October 2024.
- The project's code is hosted on a GitHub repository.[6] The repository contains the code for the chatbot and the preprocessing pipeline, together with the scripts to fine-tune the models. The `README` file contains additional information and instructions on how to run the system. Code availability will contribute to the dissemination of the project.
- A dataset consisting of the audio recorded in the videoconferences held for the pilot evaluation of the chatbot, together with their transcription, will be published under an open license. A fine-tuned error-preserving speech-to-text Whisper-based model and the corresponding training dataset will be published in the HuggingFace hub.

We plan to disseminate the project results within our university, and also to write a second paper focused on the outcomes of the human evaluation.

1.4 Ethics

Since the human evaluation involves collecting and distributing data from participants, special care has been taken to adhere to relevant ethical guidelines and applicable data protection laws. Specifically, the research ethics committee of University of Alicante has overseen the experimental process. Each participant was informed about how their interaction with the model would be used and disseminated, and they signed a consent form. Additionally, participants' personal information has been pseudonymized in the released data.

A data management plan was elaborated. It addresses issues such as data collection, data generation, data sharing, property rights and privacy, and long-term preservation and re-use, in compliance with national and EU legislation. A copy of the data management plan can be found in appendix A.

[5] https://www.youtube.com/watch?v=TzEK9JlxVH4
[6] https://github.com/transducens/demint

2 Summary of Results and Plans

2.1 Results

The development of the entire system pipeline has been completed, and the project's code has been released as open-source software. The system is capable of analyzing video conference transcriptions and providing feedback to students.

Feedback from participants in the human evaluation was gathered from two perspectives: first, the overall user experience with the chatbot, and second, the chatbot's effectiveness as an intelligent English tutor. Participants rated their responses to the evaluation questionnaire on a Likert scale from 1 to 5, with 5 representing the highest score for all aspects evaluated.

Regarding the first aspect, general user experience, participants were generally satisfied with the tool's performance and response time. In response to the question, *"Did you enjoy interacting with the chatbot?"*, all participants gave positive feedback, with a score of 4 or 5. However, fluency emerged as the system's main area requiring improvement, with a average score being 3.

In terms of the chatbot's performance as an intelligent English tutor, the overall evaluation was positive, though some areas still require enhancement. The main concern of the participants in this evaluation was the accuracy of the chatbot in identifying speech errors, which received with an average of being 3. Other aspects, such as the chatbot's ability to understand their queries, or the usefulness of examples and resources provided by the chatbot, were rated with an average score of 3.3 The clarity of the chatbot's error explanations received a slightly higher average score of 3.4. Notably, most participants agreed that the chatbot helped improve certain aspects of their English, with five out of seven giving a score of 4 for this question. Additionally, when asked whether they would be interested in using a similar chatbot in future video conferences, all participants but one gave scores of 4 or 5, demonstrating a general interest in such tools.

2.2 Business plan

Currently, there is no business plan for the project as it is in a prototype stage. In case additional funding is secured for a long-term project, a second phase of development could lead the system to a more mature stage, and a business plan could be considered.

2.3 Future plans

The system is planned to be improved via students' master theses and other local projects. Potential funding opportunities will also be explored.

Potential work for an improved version of the system includes the following:

- Supporting voice cloning to fine-tune Whisper with each student's voice before using the tool. The student will speak a few sentences and the fine-tuning data coming from C4-200M and COREFL will be used to train a customized speech-to-text model. Models such as XTTS-v2[7] could be used for this purpose.
- Considering new models such as Tower,[8] a model that also performs, among other tasks,

[7] https://huggingface.co/coqui/XTTS-v2
[8] https://unbabel.com/announcing-tower-an-open-multilingual-llm-for-translation-related-tasks

grammar error correction (GEC).

- Making the interaction with the chatbot more engaging and speech-based.
- Improving the error detection capabilities of the system, and the heuristics used to prioritize the errors to be discussed.
- Incorporating human teachers to either evaluate the error detection capabilities of the system or the interaction between chatbot and students from the point of view of the teacher.
- Integrating knowledge from theories of second language acquisition to improve the system's effectiveness.

2.4 Blurb for public dissemination on UTTER's website

DeMINT ("Automated Language Debriefing for English Learners via AI Chatbot Analysis of Meeting Transcripts") has developed a prototype of a conversational system designed to enhance non-native English speakers' language skills through post-meeting analysis of the transcriptions of video conferences in which they have participated. The code of the system is already available as open-source software on https://github.com/transducens/demint. Future plans include developing a more engaging and speech-based interaction with the chatbot and knowledge from theories of second language acquisition.

3 Recommendation by Project Sponsor

The DeMINT team has done an excellent job at delivering all results promised in the project proposal. A publication at a relevant workshop and an open-source codebase for the tool developed have been disseminated. The project has not deviated in any major way from what was proposed. The project sponsor, therefore, makes a positive payment recommendation for the DeMINT project.

A Data Management Plan

A.1 Introduction

DeMINT (Automated Language **De**briefing for English Learners via AI Chatbot Analysis of **M**eet**in**g **T**ranscripts) is a project funded in 2024 by the EU project UTTER (Unified Transcription and Translation for Extended Reality) via the Financial Support to Third Parties (FSTP) feature, also known as cascade funding. This Data Management Plan (DMP) provides an analysis of the main elements of the data management policy that have been used in DeMINT with regard to all the datasets collected for or generated by the project.

This document addresses issues such as data collection, data generation, data sharing, property rights and privacy, and long-term preservation and re-use in accordance with national and EU legislation. We describe the types of data that are collected and generated. We also look at data storage and retention, as well as data protection and the implications of the regulations to be enforced on project data, particularly with regard to the protection of personal data.

A.2 Data collected

DeMINT is developing an AI chatbot to act as a tutoring assistant for non-native English speakers. Data will be collected from both open repositories on the web with a license allowing their use for research purposes, and from user interactions through an online meeting application and through the tutoring assistant itself.

The data downloaded from open repositories will be used to train the models used by some of the individual components used by the DeMINT pipeline. They will also be used to automatically evaluate these components during their development.

Data collected from user interactions through an online meeting application will be used to conduct a human evaluation of the DeMINT tool. For this purpose, users will be recorded (both audio and video) during online meetings in which they will perform role-playing activities to mitigate the need for anonymizing the data later. After the communication is recorded, each user will follow up with a session with the chatbot to analyze the mistakes made during their meeting. Additionally, these written conversations between the user and the chatbot will also be recorded.

A.3 Data generated

We distinguish four main categories of data that will be generated during the project:

- Data generated from existing datasets to be used for training and evaluation of the individual components of the DeMINT tool.
- Audio and transcription of the online role-playing meetings.
- Debriefing text generated from the interaction with the DeMINT assistant.
- Software, models, algorithms, etc.
- Academic research publications.

A.4 Data storage, preservation and re-use

The project data will be stored in private or public repositories — depending on the privacy level of the data, see next section. The **private repositories** will reside on a machine at the Department of Software and Computing Systems of the Universitat d'Alacant and will only be accessible from the department's local network by people involved in the DeMINT project. The **public repositories** will be on GitHub, and data that is relevant will be linked from CLARIAH-ES[9] (the Spanish node of CLARIN[10]).

All data in public repositories produced during the project will be made available under free/open-source licenses.

[9] https://www.clariah.es/
[10] https://www.clarin.eu/

A.5 Privacy: levels of access and sharing

There are different categories of data collected or generated during the project, with different levels and conditions for access and sharing:

Audio and video: The raw data collected during user interactions for evaluation purposes (both audio and video) will be private. This data will be used to generate the audio and text data to be distributed under a free/open-source license. Once the data to be distributed has been generated, original data will be removed from our servers.

Identification data: This data will be associated with the audio and video collected, and will be private. It will consist of name, family name, e-mail address, gender, age, mother tongue, academic background, and level of English.

Software and models: Software, models and algorithms will be public and released under free/open-source licenses.

Debriefing text: The text of the learning interactions with the DeMINT tool will likely be made public and released under a free/open-source license. However, if it is ultimately deemed not useful for the community, it will remain private.

Data for training/testing: Data used to train or test the DeMINT tool may directly come from public repositories or be generated from other public datasets. In the latter case, we will make it public and release it under a license as open as possible, compatible with the license of the original dataset.

Academic research publications: Academic publications will be made available as "green" open-access via institutional repositories.

A.6 Personal data protection measures

This section sets out how we will identify where personal data is involved and how that personal data will be protected. Security and privacy issues will be taken into account when designing the architecture and information flow. The processing of personal data is in accordance with the data protection regulations of the Universitat d'Alacant and Spain.

- Participants will be informed of how the data will be collected, processed, and stored. To participate, they will need to be of legal age and sign an informed consent form. Identification data provided by participant (see above) will be properly safeguarded on a file in a private repository.
- To protect personal data in the data collected during user interactions (audio and video), individuals participating in such interactions (which will be recorded for evaluation purposes) will be instructed not to reveal any personal data. They will participate in role-playing activities that are unrelated to their real identities which will make the restriction considerably easier to attain.
- In case a pilot demo is put online for dissemination purposes, the interactions with users as well as the text generated by the DeMINT tool will only be temporarily stored for the purpose of functioning on the tool during the debriefing session. After that, all data will be permanently deleted.

B Feedback form filled by participants in human evaluation

General instructions: In this form, we ask you to evaluate your experience based on your interaction with the conversation bot from the DeMINT project. All questions are answered by assigning a score from 1 to 5 to assess your experience across different aspects of the bot's functionality. If any of the responses you provide are published, they will be done anonymously and aggregated with the responses of other participants in this evaluation.

In your opinion, was the conversation with the bot smooth?

Give a score from 1 to 5, with 1 meaning *Not smooth at all*, and 5 meaning *Very smooth*

Did you enjoy the experience of using a conversation bot like the one you used in this evaluation?

Give a score from 1 to 5, with 1 meaning *I didn't enjoy it at all*, and 5 meaning *I enjoyed it a lot*

Do you think the bot's response time to each of your interactions was too long?

Give a score from 1 to 5, with 1 meaning *Yes, it took too long to respond*, and 5 meaning *No, I think the response time was appropriate*

Do you believe the conversation chat accurately detected the mistakes you made when using English in the video conferences you participated in?

Give a score from 1 to 5, with 1 meaning *The error detection was terrible*, and 5 meaning *The error detection was very accurate*

Did you find the explanations provided for each mistake clear?

Give a score from 1 to 5, with 1 meaning *Not clear at all*, and 5 meaning *Very clear*

Did you find it difficult to make the bot understand your doubts regarding the detected errors?

Give a score from 1 to 5, with 1 meaning *It was very difficult*, and 5 meaning *It was very easy*

Do you think the time spent in your conversation with the bot to identify, explain, and help you improve your English based on the detected errors was excessive?

Give a score from 1 to 5, with 1 meaning *Yes, it spent too much time on the same errors*, and 5 meaning *No, the time spent on each error did not seem excessive at all*

Do you think the time spent in your conversation with the bot to identify, explain, and help you improve your English based on the detected errors was insufficient?

Give a score from 1 to 5, with 1 meaning *Yes, it spent too little time on some errors*, and 5 meaning *No, the time spent on each error did not seem insufficient at all*

Did you find the resources (examples, exercises, etc.) provided by the bot to help you improve your knowledge of English related to the mistakes you made useful?

Give a score from 1 to 5, with 1 meaning *No, the resources were not useful at all*, and 5 meaning *Yes, the resources were very useful*

Do you think that interacting with the conversation bot has helped you improve your spoken English in any way?

Give a score from 1 to 5, with 1 meaning *I don't think it helped me at all*, and 5 meaning *Yes, I think some of the corrections or suggestions it provided helped me improve*

Do you think more sessions with a conversation bot like this one (in future improved versions) could help you improve your English?

Give a score from 1 to 5, with 1 meaning *I don't think it would help me at all*, and 5 meaning *I think it would help me a lot*

Would you use a conversation bot like this again (in future improved versions) to improve your English after participating in English video conferences?

Give a score from 1 to 5, with 1 meaning *No, I don't think I'd like to use it again*, and 5 meaning *Yes, I'd love to use it after each video conference*

B.6 SURE-GB

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP1 Evaluation – SURE-GB

Identifying Stereotypical, Under-representational, and Algorithmic Gender Bias in Machine Translation

Nature	Evaluation Report	**Work Package**	WP1
Project start date	dd/mm/2024	**Project end date**	dd/mm/2024
Interim meeting	dd/mm/2024	**Report submission Date**	dd/mm/2024
Main authors	Chrysoula Zerva & Ben Peters (IT)		
Reviewers	Maryam Hashemi (UVA)		
Version Control			
v0.1	**Status**	Draft	XXX
v1.0	**Status**	Final	XXX

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Evaluation of Project Execution

The **SURE-GB (Stereotypical, Under-representational, and Algorithmic Gender Bias in Machine Translation)** project aimed to develop an automated system to detect and mitigate gender bias in machine translation (MT) system outputs, focusing on occupation-related language across English, French, and Greek. The project addressed gender bias by developing a curated Knowledge Graph (KG) that integrates real-world gender occupation statistics from official sources at EU and national levels as well as linguistic features from large corpora. The SURE-GB team aims to employ this KG to identify different sources of gender bias in occupational terms.

The evaluation of the project in terms of the submitted deliverables, published work, and online meetings and presentations, revealed several strengths, some challenges, and a great potential for future expansion and exploitation of the released resources.

1.1 Project Execution and Achievement of Milestones

The project successfully met key milestones, ensuring timely completion of deliverables. The core milestones, including data collection, KG development, bias detection system implementation, and dissemination, were achieved with only minor deviations.

Milestones 1 and 3: Data Collection and Knowledge Graph Development The project team constructed a **Knowledge Graph** based on standardized data, such as EU-LFS and ISCO-08, encoding occupation-related gender statistics for English, French, and Greek texts. The KG structure integrates employment data from Greece, France, and the United Kingdom, alongside language-specific textual corpora which were selected based on popularity of usage for testing and training MT models. This resource enables analysis of gender biases at different occupational levels, providing a detailed representation of how occupations are "gendered" in the labour market versus textual data used for MT. There were no deviations in the promised implementations and both collected data and the KG are available on the team's GitHub page with the provided code to reproduce the KG with different data as well.

Milestone 2: Gender Mismatch Detection Module The **gender mismatch detection tool**, developed as part of the project, successfully identified disagreements in gender assignment between source and target texts during MT. The team explored several alternatives (including some suggested in the interim meetings) and chose to use a hybrid method that relied on large language models (LLMs) for the detection of terms of interest. They evaluated on known corpora and showed their method to have high accuracy when using open-access LLM variants (Llama 2) that are on the "larger" side (70B parameters). The rest of the mismatch detection pipeline relied on traditional NLP modules (i.e. the SpaCy library) for gender detection.

Milestone 4: Gender Bias Detection System The team built a fully operational **bias detection system** that leverages the KG to classify gender biases in MT outputs. By analyzing gender representations within official statistics and textual datasets, the system was able to detect three types of bias: *under-representational*, *stereotypical*, and *algorithmic* bias. Hence, the proposed method can uncover significant flaws in existing MT systems, particularly in cases where the biases are misaligned with real-world occupation statistics (i.e. in cases of stereotypical or algorithmic

bias). The authors demonstrated their system on the UTTER user day, and the implementation is accessible on the team's GitHub page.

Milestone 5: API Integration An **API endpoint** was developed to facilitate the integration of the gender bias detection tool into other systems. Although the core functionality was delivered, it has been noted that there is no guarantee for the maintenance and continued deployment of the API upon completion of the service, due to additional costs that would need to be covered. Although this differs from the plan, it does not constitute a significant deviation because the source code for the implemented tools is available and the API is documented.

Milestone 6: Dissemination The project successfully disseminated its findings through academic publications, open-source repositories, and a public video presentation, with plans for additional publications.

1.2 Key Achievements, Strengths and Challenges

Overall, the presented work constitutes a significant interdisciplinary contribution that bridges knowledge from the social sciences and real-world occupational analytics with statistics from textual data used for training LLMs. The proposed method provides a novel way to "decompose" bias observed in the output of LLMs used for MT and could help better understand and mitigate biased text generation. As a further point for improvement, it would be interesting to consider whether there are additional resources that could be used to understand the sources of algorithmic bias, i.e. bias that cannot be attributed to imbalances found in either the occupational or the textual data.

Additionally to the automated evaluation of existing benchmarks, the authors are carrying out a human evaluation campaign to further validate their methods and obtain a relevant dataset, which could provide a useful resource for evaluating future models and methods. However, the team noted that the annotation process is demanding because it requires annotators to be familiar with EU standards for occupational classification (ISCO standards); annotators with the relevant expertise are difficult to find.

It should be noted that although French and Greek use different scripts and come from different Indo-European subfamilies (Romance and Hellenic, respectively), they provide a limited picture of the diversity of grammatical gender phenomena in the world's languages. It would be great to see expansion of this work into further languages in terms of both presented analysis and human annotations.

Overall, the project demonstrated an interesting perspective on addressing gender bias in MT and language generation systems and laid the groundwork for future research in this critical area.

1.3 Overall Recommendation by Project Sponsor

Given the clear achievements, adherence to the proposal, and the potential for future impact and expansion, The SURE-GB team has shown more than satisfactory performance and already produced several outputs. Hence, it is recommended that the SURE-GB project receives the final funding part as originally planned.

B.7 InCroMin

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP1 Final – InCroMin

Interactive Crosslingual Minuting

Nature	Final Report	**Work Package**	WP1
Project start date	01/01/2024	**Project end date**	30/09/2024
Interim meeting	26/04/2024	**Report submission Date**	30/09/2024
Main authors	XXX Sponsor (PARTNER)		
Co-authors	Ondřej Bojar, Marko Čechovič, Natália Komorníková, Dominik Macháček, Peter Polák (CUNI)		
Reviewers	Maryam Hashemi (UVA)		
Version Control			
v0.1	**Status**	Draft	–
v1.0	**Status**	Final	30/09/2024

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Project Execution

1.1 Deviations from original plan

The planned goals of InCroMin were:

1a) To further develop MinuteMan by merging it with UTTER components.
Deviation: We consulted MinuteMan integration options with the UTTER team. We agreed on not merging MinuteMan with UTTER components because at that point, there was not a single UTTER pipeline. MinuteMan can be regarded as a standalone item in the collection of UTTER tools.

1b) To expand MinuteMan for the multi-lingual meeting situation.

No deviation.

2) Continue our research on multi-source speech transcription and translation, in order to benefit from human interpretation live.

Deviation: While interpreters now regularly serve in remote calls, it would be substantially more difficult and expensive to obtain examples of such calls. We therefore shifted our research focus on a different challenge: automatic interpretation from sign language.

3) Collect a test set of naturally multilingual meetings (harder to get access to), or as a fall-back monolingual meetings (easier to get access to and also already partially available to us) with cross-language access needs to them.

No deviation.

4) Evaluate multi-lingual interactive meeting summarization in practice and on the test set.

Deviation: Early test calls in the MinuteMan platform indicated that live meeting summarization quality is insufficient. We thus focused on the first step only: evaluating the cross-lingual meetings themselves, not their summaries. We nevertheless prepared two datasets (test-set size) for cross-lingual meeting summarization: many of our test calls have manually created minutes by our annotators (Appendix A.2) and we translated ELITR-Bench into Czech (see Objective 4 below).

Planned InCroMin project outputs:

1) Interactive meeting summarization tools (based on MinuteMan) extended for multilingual use.

As decribed in Goal 4 above, we limited our attention to summarization and instead focussed on the multilingual use of MinuteMan and the effectiveness of cross-lingual meetings as such.

2) A curated corpus of meetings, usable for evaluation and testing of speech translation and meeting summarization into minutes. Similar to ELITR Minuting Corpus but with multi-lingual meetings.

No deviation.

3) **Speech transcription and translation models modified for multi-sourcing, so that they can benefit from human interpretation.**

   Deviation: Instead, we deliver progress in sign language translation research.

Planned dissemination:

- **For academic and educational sector.**

  Negligible deviation: we stayed at university level of education. We tested cross-lingual MinuteMan use for academic and education use in several calls (consultations on both content of study as well as life and organizational matters) but we did not test the tool with any high-school level institutions.

- **For citizen support sector.**

  No deviation. Our test calls include professionals from the Integration Center Prague (ICP), an NGO focused on citizen support.

- **Publish the results of our research in the relevant top-tier peer-reviewed research conferences or journals.**

  Pending. While we have gathered enough content, the 9-month project time span proved too short to be able to polish it into a paper submission.

1.2 Development

Objective 1. Facilitate cross-lingual access to meeting transcripts/translations.

Multilinguality in MinuteMan Before InCroMin, MinuteMan (Kmječ and Bojar, 2023) was an application for real-time meeting transcription and summarization that supported only one language – English. Within InCroMin, we added the multilingual support, to enable cross-lingual communication between meeting partners who do not share a common language. We carried out the following steps:

- **Software Engineering** – MinuteMan required significant software engineering improvements, such as better logging, more stable deployment, and resolving data persistence issues. Next, the user interface was simplified based on initial users' feedback. After following these upgrades, multilingual transcriptions and summaries for cross-lingual meetings could have been introduced.

- **Multilingual models** – We integrated multilingual automatic speech recognition Whisper large-v3 (Radford et al., 2022), and a multi-lingual MT model NLLB (Costa-jussà et al., 2022) that is capable of translating English into 200 other languages. Implementation of these two models went quite smoothly, but some parameter tuning needed to be done to achieve suitable results.

- **Automatic data collection** system was implemented the collection of InCroMin Test Calls corpus (ref. Objective 3).

During the development of these features, challenges were encountered with the application design, particularly regarding extensibility, as well as user-technical difficulties during testing. Most issues were resolved, but some questions remain, particularly concerning audio recording buzzing, which may be caused by inconsistencies in the online web calling platform clients and were sometimes resolved by simply reconnecting the participants.

Based on the data and experience collected, the following further improvements are proposed: (1) Better voice activity detection to avoid translating noise into non-sense sentences, or cutting off unfinished sentences too aggressively. (2) More complex machine translation (MT) pipeline for specific language combinations, such as direct Czech-Ukrainian MT (Popel et al., 2024), could help with the translation quality. (3) Users' feedback-driven development of a user interface for asking questions about meetings would be beneficial. The last mentioned suggestion is in line with the idea of ELITR-Bench test set, see page 10.

Sign Language Translation Using a sign language or signs for communication is natural for many people, including but not limited to deaf and hard of hearing and their family members. In USA, there are 9 millions of people who report using signs in any period of their lives (Mitchell and Young, 2022). In other countries, the same proportion of 2.8% of sign language users in the population is assumed. In anyway, sign language users would largely benefit of being included into cross-lingual communication thanks to machine translation support, so the sign languages are very relevant for language technology providers. The problem is that the machine translation of sign languages is relatively underdeveloped. The state of the art, such as Zhang et al. (2024), require large computational and data resources, and still report results practically unusable in real-life application. Therefore, we aimed to focus on sign language translation research.

For that, we joined an international and interdisciplinary team of researchers at the intensive research workshop JSALT 2024.[1] The team worked on the research prototype "SignLLaVA: Sign Large Language and Visual Assistant." It is based on the general concept of LLaVa (Liu et al., 2023a,b). SignLLaVa inserts sign language video features converted through projector layer into the common vector space of the language tokens of the Llama large language model (Dubey et al., 2024). SignLLaVa aims to translate American Sign Language (ASL) into English and serve as an assistant that can e.g. answer questions about the sign language video.

The unique contribution of SignLLaVa includes applying three complementary sign language representations, a demonstration tool, and an effort to create new authentic ASL-to-English MT test set, which would resolve the problems of currently common state-of-the-art How2Sign data set (Duarte et al., 2021), which is created for the opposite direction English-to-ASL. There is an unrealistic optimal segmentation into sentences, and other problems. The progress in the ASL-to-English test set includes a plan to identify suitable ASL videos that have English translations and are available on the Internet, and a small initial probe with several videos.

InCroMin team members complemented the SignLLaVa team by the necessary expertise in machine translation, including the MT evaluation and creating the test set. Moreover, they contributed by e.g. text data normalization, and application of LLMs to texts for multi-tasking and evaluating the sign language representations. The SignLLaVa team did a significant progress towards application of sign language translation into MinuteMan and other cross-lingual communication tools.

[1] https://www.clsp.jhu.edu/sign-language-translation/

Whisper-Streaming improvements Whisper-Streaming (Macháček et al., 2023) is a tool for real-time transcription and translation of 99 source languages that are supported by Whisper. It is a necessary component of cross-lingual meeting tools. In InCroMin, we focused on technical improvements of Whisper-Streaming, to enable its integration into MinuteMan and other similar tools. Within InCroMin, we added the following features:

- **Automatic language identification**, the same method that is implemented in Whisper. The language identification is applied on every update, so it allows fully automatic switching between the source languages.
- **Voice Activity Controller** using Silero VAD Iterator. Whenever there is e.g. 0.04-second audio chunk, we run an iterative Voice Activity Detection (VAD) model to detect beginning or end of speech that has to be processed. It improved quality by avoiding Whisper to process non-voiced segments, which often lead to hallucinations, and improved latency because the end of speech (a significant pause, 0.5 seconds) triggers immediate update, not waiting for the next chunk for confirmation.
- **OpenAI API backend**. A new alternative backend that does not require local hardware for deploying the Whisper model was proposed and implemented by one GitHub contributor in our cooperation. Our tests showed that processing trough API achieves the same quality as local processing, but the latency is much larger and unpredictable. There is also a significant cost for the API, with 1 second audio chunk approximately 10-times higher than processing the offline audio once. Moreover, the API seemed to be changed in the newer version and following code maintenance is needed. The faster-whisper backend with local deployment of Whisper model is still the most recommendable option, but the API showed an alternative way that may be useful for some applications.

There is relatively large and active community of users and developers of Whisper-Streaming on the open-source code repository GitHub, documented by nearly 1 800 stars and 200 forks in September 2024.[2] Within InCroMin, we cooperated with them, responding to their issues and pull requests. Within InCroMin period—between 01/01/2024 and 24/09/2024—we responded to 77 issues or pull requests. Although many of them were relatively simple clarification questions, issues with installation, usage or quality, several issues or pull requests led to large valuable improvements, such as Voice Activity Controller, and some others to small but useful improvements such as removing duplicated variable, improved debug logging, adding warmup file to make Whisper-Streaming server to process the very first audio chunk faster, etc.

In summary, we were managing and supporting the open source community around Whisper-Streaming, and we gained lots of benefits that contributed to the InCroMin goals.

Objective 2: Facilitate cross-lingual access to meeting minutes.

The original goal of InCroMin was to provide participants of the meetings with live summary of the meeting in their language. This goal was unfortunately too ambitious for the short time span of the project. We nevertheless conducted initial experiments with LLMs for summarization. We used LLMs with 7 billion parameters (specifically this mlabonne/NeuralBeagle14-7B)[3] to obtain

[2] https://github.com/ufal/whisper_streaming
[3] https://huggingface.co/mlabonne/NeuralBeagle14-7B

summarizations, showing promising results that could be tailored to user's preference by adjusting the initial prompt. However, due to the complexity of the MinuteMan's architecture, these enhancements were not integrated. The LLMs also demonstrated the ability to reliably answer questions about the meetings, which could be particularly useful for future references. For the time being, we did not evaluate this step yet and only managed to prepare the relevant test set, see ELITR-Bench Czech on page 10.

Objective 3: Prepare test sets for cross-lingual meetings.

While low-latency speech translation systems are publicly available and sometimes even built to remote conferencing platforms (e.g. Zoom) or to cell phones (Samsung Live Translate),[4] we believe that the true level of "penetration" of the language barrier when making a cross-lingual call has not been properly assessed yet.

To this end, we organized calls between participants with no mutual language understanding, so they have to rely on speech-to-text translation provided by our MinuteMan. The calls are topically diverse, spanning from travel or living-abroad experience up to regular examples of project meetings or technical consultations. We collected the sound and transcripts of the calls, we manually revised them, and organized them in the **InCroMin Test Calls** corpus. We also collected experience of the participants in a feedback questionnaire. To complement the subjective assessment, and to start a semi-formal analysis of misunderstandings in communication, we equipped the corpus with an annotation of misunderstandings. Through this we hope to get a better idea of how frequent misunderstanding are, how often they can be attributed to speech translation errors, what needs to be fixed first to reduce misunderstandings count, and also if large language models are capable of identifying misunderstandings in meeting transcripts.

Details on this activity are provided in the report in Appendix A

Objective 4: Rigorous evaluation of underlying models.

Speech translation support for cross-lingual calls can be realized using a considerable number of architectures and technical components. Some setups can be end-to-end, with one model achieving the full needed process, some setups can first transcribe the sound and then translate the transcribed text to the target language.

Thanks to InCroMin, we contribute to rigorous evaluation of the necessary components on two fronts:

- **Analysis of automatic evaluation of speech translation latency**. A crucial aspect of a speech translation system for cross-lingual meetings is its latency, i.e., the duration required to provide the translation to the users. Unlike in speech recognition, latency measurement in translation is complex due to reordering. In recent years, researchers have proposed various latency measures. Currently, the IWSLT campaign (Ahmad et al., 2024) employs five different latency metrics to evaluate submitted systems. However, our detailed analysis in Appendix F reveals that these metrics do not correlate strongly with each other. To improve

[4] https://www.samsung.com/latin_en/support/mobile-devices/how-to-use-live-translate-for-phone-calls-on-the-galaxy-s24/

the reliability of InCroMin system evaluations, we examine which metrics reflect actual latency with the highest precision.

Results show that the DAL (Arivazhagan et al., 2019) is the most robust metric and should be used for comparing different systems. Alternative latency metrics can be employed in system development, as they generally show a good correlation with true latency when the compared systems are similar.

More details are in Appendix F.

- **Dialogue Translation Test Sets** Dialogue and conversational content in general are not yet well covered in established evaluation campaigns such as WMT or IWSLT, although they bring specific challenges not seen e.g. in news text or monologues. We used InCroMin funds to prepare test sets geared towards InCroMin goals for recent as well as future use: WMT24 evaluations, MultiWOZ Czech and German Small Test Set, and ELITR-Bench Czech.

WMT24 Translation Evaluation The Czech-to-Ukrainian WMT24 test set source was collected through the Charles Translator for Ukraine[5] (Popel et al., 2024). With users' consent, the service can log their inputs for the purpose of creating a dataset of real use cases. The data includes cross-lingual dialogues of Czechs communicating with Ukrainians (mostly refugees). There is also a subset of originally spoken data from users using the Charles Translator mobile app. The Czech source translations were translated to Ukrainian by professional translators (funded from other sources) in order to create the gold reference for WMT 2024[6] and InCroMin funds were used for the manual evaluation of this language pair and also English-to-Czech. The results will appear in WMT24 Findings.

MultiWOZ Czech and German Small Test Set For the purposes of evaluation of automatic translation of meeting transcripts, we had professionally translated a small portion of the MultiWOZ (Ye et al., 2022)[7] dialogues dataset from English into Czech, and as a second step from Czech into German. This two-step procedure was adopted to maintain consistency in important linguistic features which are often not explicit in English but which are expressed in Czech and German, primarily the gender and level of politeness. We instructed the translators to attribute gender to the speakers arbitrarily and ensure consistency within each dialogue.

The actual use of these test sets for MT system development or selection was not carried out in the limited time of InCroMin. Instead, we will include these test sets into WMT25 General Translation Task and possibly also WMT25 Test Suites, where we would be assessing phenomena critical for dialogue fluency such as gender and politeness preservation.

Interleaving our translated dialogues with each other or with the original English version allows us to construct also simulated cross-lingual setting and develop multilingual access methods for this content. This is however left for future work.

MultiWOZ Czech and German Small Test Set is available upon request from Ondřej Bojar, until it will have served in WMT evaluations.

[5] https://translator.cuni.cz/
[6] https://www.statmt.org/wmt24
[7] https://github.com/smartyfh/MultiWOZ2.4

Q: Who were the participants of the meeting?
A: [PERSON14], [PERSON10], [PERSON5], [PERSON9], [PERSON1], [PERSON11]
Q: What was the main purpose of this meeting?
A: Discuss and finalize the technical setup for a demo
Q: How many scenarios were discussed?
CONTEXT-FREE Q: How many scenarios were discussed for the demo setup?
A: 3 (plans A, B and C)
Q: Which scenario was chosen eventually?
CONTEXT-FREE Q: Which scenario was chosen eventually for the demo setup?
A: Plan C

Figure 1: An illustration of context-sensitive and context-free variants of ELITR-Bench questions as provided to translators.

ELITR-Bench Czech ELITR-Bench (Thonet et al., 2024) is a collection of questions and answers in English created to complement English meetings from the ELITR Minuting Corpus (Nedoluzhko et al., 2022) with questions and golden-truth answers. This test set is meant to evaluate the accessibility of information in meeting minutes with QA systems or LLMs.

In InCroMin, we had ELITR-Bench questions professionally translated from English into Czech. This variant of the test set allows anyone to evaluate *cross-lingual* access to meeting content: asking questions in Czech, locating answers in English meeting minutes, reporting answers in Czech and comparing them to the golden-truth Czech answer.

ELITR-Bench contains questions in two settings: as a continuous dialogue where the formulation of the question may need the previous context of the dialogue, and as independently formulated questions. When preparing the translation batches for translators, we noticed that these two settings overlap greatly. Of the total of 271 question we were translating, only 33 have a separate context-free formulation. We substantially saved the translation costs by providing translators with the 271+33 questions in a single sequence, with context-free variant coming right after the context-dependent one, as illustrated in Figure 1.

ELITR-Bench Czech is available upon request from Ondřej Bojar, to prevent LLMs learning from it accidentally.

1.3 Dissemination

Research seminars InCroMin was twice briefly presented as an ongoing project at research seminars within one-hour lecture of Dominik Macháček who presented his PhD. research "Multi-Source Simultaneous Speech Translation." First, at the Linguistics Mondays at ÚFAL MFF CUNI (Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics) on 04/03/2024.[8] There were around 20 participants on-site and online. The lecture recording on the institute's website[9] has 387 views, YouTube[10] reports 25 views and 318 subscribers.

Second, at the research seminar of the CSTR and ILCC groups at the University of Edinburgh, School of Informatics, on 22/04/2024. There were approximately 20 attendees.

[8] https://ufal.mff.cuni.cz/events/multi-source-simultaneous-speech-translation
[9] https://lectures.ms.mff.cuni.cz/view.php?rec=534
[10] https://www.youtube.com/watch?v=FpumkKjCJO0

JSALT 2024 Closing Day The SignLLaVa team presented their results at JSALT 2024 Closing Day on 02/08/2024. There were approximately 65 attendees on-site, the YouTube recording[11] has 566 views and 2 380 followers.

Machine Translation Marathon is a week-long gathering of machine translation researchers, developers, students and users. In 9/2024, it was organized at ÚFAL MFF CUNI[12] in Prague, Czech Republic, and there were approximately 50 participants. InCroMin was presented to the participants at several points:

- Project proposals,
- Project midweek and final reports,
- Whisper-Streaming demo at the open poster session.[13]

The collection of slides presented at MT Marathon is provided in Appendix C.

UTTER Users Days We presentated InCroMin project at UTTER Users Days online conference, highlighting the goals, some technical details, and progress of this project. Hopefully we were inspiring enough for other attendees to consider FSTP project funding as well. The slides are provided in Appendix D.

ELF Conference Ondřej Bojar presented InCroMin at the English as Lingua Franca international conference[14] organized by the Prague City University. The presentation raised attention because relying on speech translation goes against the spirit of the universal and inclusive use of English. It is however clear that the best approach is to combine InCroMin-like tools (in early stages, when mutual understanding is not possible) with gradual adoption of a common language such as English. As a follow-up of this dissemination, one test call was acquired and there are good connections established between us and some ELF participants for future joint research where the new colleagues would cover primarily "soft" aspects of the task, e.g. running InCroMin test calls with non-technical people, or evaluating the (mis-)understandings more rigorously.
The slides are provided in Appendix E.

In-house presentations We made use of the opportunity of ÚFAL offsite seminar to informally introduce InCroMin to our colleagues and students from our department we don't often get in touch with. As a result, several InCroMin test calls were again acquired. We also showcased the project to our colleagues in Gen Digital Inc. which sparked their interest and also motivated them to voluntarily provide us more test calls and valuable feedback. They were happy to assist with any further testing of similar applications.

1.4 Ethics

In InCroMin, we followed our established practice of consent collection. The practice was defined during the EU project ELITR and conforms to the standards of Charles University.

The consent form was updated to reflect InCroMin test calls corpus collection. See Appendix B for the full text.

[11] https://www.youtube.com/watch?v=65L7tklQbyc
[12] https://ufal.mff.cuni.cz/mtm24/
[13] https://ufal.mff.cuni.cz/mtm24/abstracts.html
[14] https://www.praguecityuniversity.cz/elf

The data collected into the InCroMin corpus were deidentified by removing all personal names from the transcripts and the subsequent documents (translations and summaries). The occurrences of participants' names were also silenced in the provided recordings.

Depending on the circumstances, the participants were paid or unpaid volunteers. We paid participants from the Integration Center Prague (IC Praha) and several other participants with no direct interest in language and speech technologies. Another set of volunteers was solicited from students of relevant subjects at Charles University and from participants of the JSALT 2024 workshop; these participants donated their time for free because they were interested in testing out the state-of-the-art system from their field of study. The last group were participants from MT Marathon. It was technically impossible to pay these participants (our university would require to prepare short-term contracts with them, which is simply infeasible during the one week of MT Marathon), so we provided them with a souvenir from Prague.

2 Summary of Results and Plans

2.1 Results

This is the summary of tangible outputs of InCroMin.

- Cross-lingual meetings tools:
  - **MinuteMan**: The project is fully open-sourced and well documented at GitHub page https://github.com/fkmjec/minuteman.
  - **Whisper-Streaming**: The commits and author's activity in the public repository https://github.com/ufal/whisper_streaming are outputs of InCroMin.
- Research results:
  - **Sign language translation**: The output is research progress documented in JSALT 2024 Sign LLM, Large Sign Language Model team final report.
  - **Analysis of latency measures for speech translation**: The output is technical report in Appendix F.
- Datasets:
  - **InCroMin Test Calls** is described here in Appendix A and the deidentified data are publicly available at https://github.com/ELITR/incromin-test-calls
  - **MultiWOZ Czech and German Small Test Set** is available upon request from Ondřej Bojar, until it will have served in WMT 2025 evaluations.
  - **ELITR-Bench Czech**, the translation of ELITR-Bench into Czech, is available upon request from Ondřej Bojar; aimed for future evaluations of LLM applicability in cross-lingual access to meetings.

2.2 Business plan

We do not have any business plan that would directly exploit InCroMin results.

2.3 Future plans

- **Research publications**. We plan to publish research publications with the following content:
  - Sign language translation research results.
  - Analysis of lantecy measures in speech translation.
  - InCroMin Test Calls, including misunderstanding annotation.
  - ELITR-Bench Czech.
  - MultiWOZ Czech and German Small Test Set.
- **Research projects**: We used InCroMin findings and preliminary results for prioritization and motivation for proposing future research projects, including but not limited to Horizon Europe MSCA postdoctoral fellowship project on live credible translation.
- **Student projects**: We propose projects and theses to Charles University students, such as cross-lingual communication tool with synchronization, multi-lingual post-editing, etc. A reimplementation of MinuteMan is proposed as a team software project.
- **MinuteMan**: Future plans with MinuteMan are: (1) Integrate Whisper-Streaming with enhanced voice activity detection as a module. (2) Finish the implementation of propagation of user edits of transcripts to other languages. (3) Optional automatic transcript scrolling. (4) Explore the idea of generating subtitles from the transcripts in real time so it could be embedded into a remote call software removing the need of external application for users that only need to watch and not interact with the transcripts.

2.4 Blurb for public dissemination on UTTER's website

In InCroMin, we examined and carefully evaluated the applicability of recent state-of-the-art speech-to-text translation tools in real cross-lingual calls, i.e. calls between parties that do not have a common language. We adapted MinuteMan (https://github.com/fkmjec/minuteman) for this purpose and collected a corpus of such calls. The deidentified part of the corpus is available here: https://github.com/ELITR/incromin-test-calls. Additional results of InCroMin include an evaluation of latency metrics for speech translation, translation of ELITR-Bench (https://github.com/utter-project/ELITR-Bench) into Czech to allow evaluation of cross-lingual access to past meeting content or translation of a part of MultiWOZ dialogues into Czech and German to assess translation quality of dialog-critical features such as participants' gender preservation. All the outputs are detailed in InCroMin Final Report.

3 Recommendation by Project Sponsor

From:	Laurent BESACIER
To:	Maryam Hashemi Shabestari
Cc:	Barry Haddow
Subject:	FW: InCroMin Final Report
Date:	woensdag 2 oktober 2024 18:04:45
Attachments:	SUBMITTED.pdf InCroMin Wrap-up Meeting Slides.pdf

---

Dear Maryam

Here is the InCroMin final report they submitted
We also had today the final wrap up call during which they shared some slides (also attached to this message)

Based on these two documents and on our discussions, here is our final assessment (which is positive for unlocking the 2d part of the project money) - they did not share the overleaf so we could not include it directly into a single report, sorry for that.

=======
We had a very productive review meeting for InCroMin. Overall, the project exceeded expectations. In summary, they:
- Adapted MinuteMan to support cross-lingual calls.
- Collected a new and potentially valuable corpus of simulated cross-lingual meetings.
- Conducted practical tests to assess the usability of the extended MinuteMan and identified areas for improvement.
For well-supported languages, MinuteMan appears close to being fully operational. Additionally, the potential founding of a spin-off for MinuteMan is under consideration, with FSTP funding playing a crucial role in bringing the system closer to production-ready.
Finally, even though it wasn't initially planned, InCroMin developed a Czech version of the ELITR-Bench meeting, which will soon be added to the UTTER/ELITR-Bench repository. This could also spark future collaboration between Naver and Charles University on cross-lingual Q&A on long documents (meeting transcripts).
======

Best
Laurent & Barry

-----Original Message-----
From: "Ondrej Bojar"<bojar@ufal.mff.cuni.cz>
To: "Laurent BESACIER"<laurent.besacier@naverlabs.com>; "Barry Haddow" <bhaddow@staffmail.ed.ac.uk>;
Cc: "Dominik Machacek"<machacek@ufal.mff.cuni.cz>; "Marko Cechovic" <markocechovic@gmail.com>; "naty komorka"<naty.komorka@gmail.com>; "Peter Polak, FW" <polak@ufal.mff.cuni.cz>;
Sent: Mon, Sep 30, 2024 22:29 (GMT+02:00)
Subject: InCroMin Final Report

Dear Laurent, Barry,

attached please find our final report for InCroMin.

I am not sure what the "Sponsor (PARTNER)" cell should be, so I put all of us authors into Co-authors.

Talk to you on Wednesday at 14.00 Prague time.

Thanks,

Ondrej.

--
Ondrej Bojar (mailto:obo@cuni.cz / bojar@ufal.mff.cuni.cz)
http://www.cuni.cz/~obo

B.8 pyannote.mobile

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP1 Final – pyannote.mobile

On-device streaming speaker diarization

Nature	Final Report	**Work Package**	WP1
Project start date	15/01/2024	**Project end date**	14/10/2024
Interim meeting	24/06/2024	**Report submission Date**	30/09/2024
Main authors	Marcely Zanon Boito, Laurent Besacier (NAVER LABS)		
Co-authors	Hervé Bredin (CNRS, IRIT)		
Reviewers	Maryam Hashemi (UVA)		
Version Control			
v1.0	**Status**	Final	30/09/2024

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Project Execution

1.1 Deviations from original plan

No major deviation from the original plan.

1.2 Development

Objective 1. Streaming extension of `pyannote.audio` open-source toolkit

Support for streaming speaker diarization has been added to `pyannote` toolkit (Bredin (2023)).

The offline speaker segmentation model architecture (Bredin and Laurent (2021); Plaquet and Bredin (2023)) has been adapted to work in causal manner with support for variable latency (between 0ms to 1s). Changes include removing instance normalization step, switching from bi-directional to unidirectional internal recurrent neural networks, and adding a look-ahead mechanism. The inference pipeline has also been adapted to support this new type of causal segmentation model. This allowed us to dive deeper and more efficiently into the study of the latency/accuracy trade-off. The work achieved in this part of the project has been published at Interspeech 2024 (Rahou and Bredin (2024)). Paper abstract is repeated here for convenience:

> *We address the task of streaming speaker diarization and propose several contributions to achieve a better trade-off between latency and accuracy. First, computational latency is reduced to its bare minimum by switching to a causal frame-wise speaker segmentation architecture. Then, a multi-latency look-ahead mechanism is used during training to support adaptive latency during inference at no additional computational cost. Finally, we detail the method used during inference to achieve the final frame-wise segmentation. We evaluate the impact of these contributions on the AMI meeting dataset with a focus on the speaker segmentation step, seen through the prism of voice activity detection, overlapped speech detection and speaker change detection.*

Objective 2. Proof-of-concept of on-device streaming speaker diarization

A `pyannote` SDK targeted at iOS and macOS platforms has been developed and provides a fully end-to-end streaming speaker diarization API to be used in iOS and macOS apps.

Two demo apps depicted in Figure 1 (one on iOS and one on macOS) have also been developed to showcase how developers can use this new streaming SDK. Main features include live speaker diarization from the microphone (smartphone or laptop), batch speaker diarization from an existing recording, replay/visualization of the results, and export to CSV file format.

1.3 Dissemination

Research results obtained in *Objective 1* have been presented to the speech processing community at Interspeech 2024 in September 2024 in Kos (Greece), as a poster repeated in Figure 2 for convenience.

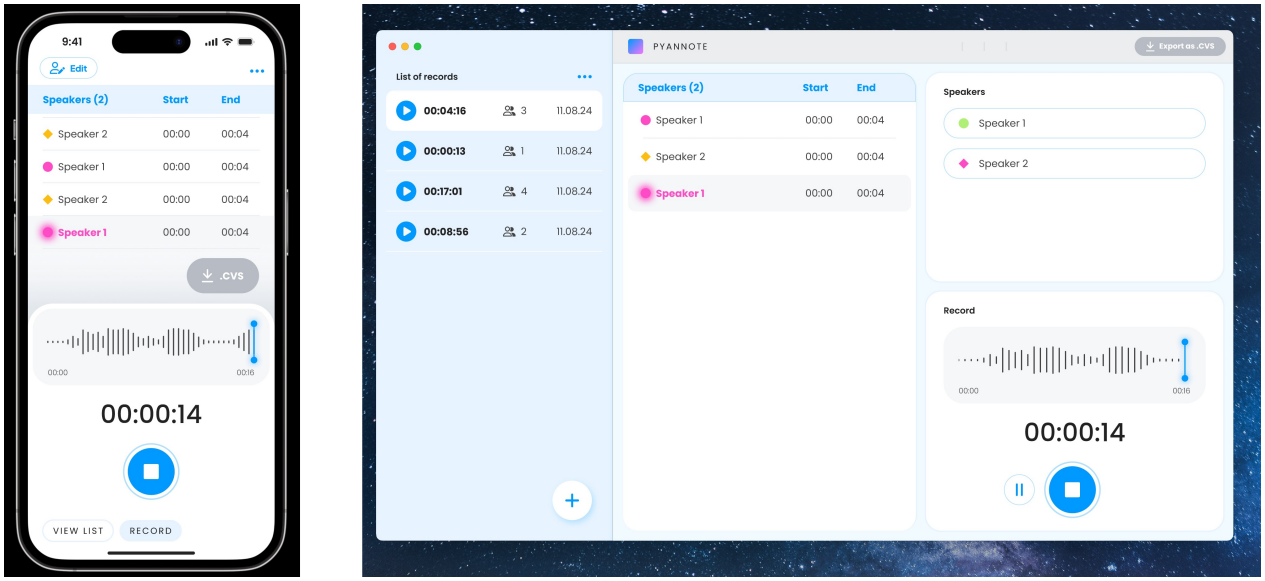


Figure 1: Screenshot of iOS and macOS demo apps

Streaming support has been added to `pyannote.audio` open-source toolkit. MIT-licensed code is currently being reviewed, already functional and available for anyone to try at the following address: `github.com/pyannote/pyannote-audio/pull/1544`.

Finally, beta builds of iOS and macOS demo apps are being uploaded to `Apple TestFlight` and we started granting access to a first batch of beta testers (mostly iOS app independent developers) for feedback.

iOS and macOS streaming speaker diarization SDKs are implemented in `Swift` and are not meant to be open-sourced in the short term as there are plans to commercialize them (see next section).

1.4 Ethics

This project dealt with the automatic processing of live recordings of conversations between several people and relies on deep learning models, that relies themselves on large collection of data that are known to contain societal biases.

We did not record any new personal data for the purpose of this project. We relied exclusively on existing and well-established academic speaker diarization benchmarks (such as AMI, VoxConverse, DIHARD, or AliMeeting).

Though the proposed technology can be used to help people better communicate (e.g. for deaf or hard-of-hearing people), we are also well aware that such speaker recognition technologies, if used by the wrong people, may lead to less desirable applications such as mass-surveillance for instance. We do believe, however, that open-sourcing the technology allows anyone to study and understand it, and therefore raise awareness of the general public.

2 Summary of Results and Plans

2.1 Results

We extended `pyannote.audio` open-source speaker diarization toolkit by adding support for the streaming scenario (where audio streams are processed in real-time instead of in batch after they completed).

We also developed a proof-of-concept of on-device (iOS/macOS) streaming speaker diarization, including an SDK implemented in Swift that we plan to distribute to interested actors in the field, through the local university tech transfer office.

2.2 Business plan

As stated above, we will work hand-in-hand with the local university tech transfer office (Toulouse Tech Transfer) in order to look for potential industrial partners interested in the iOS and macOS SDKs. In particular, `pyannoteAI` (a company building on top of `pyannote` open-source toolkit, co-founded by Hervé Bredin) will likely become one of the first user of this new piece of technology.

2.3 Future plans

Future plans include a collaboration with `pyannoteAI` university spin-off company, for them to distribute the real-time speaker diarization SDK to interested companies building iOS and/or macOS apps. Extension to other platforms such as Android phones, Raspberry Pi, or even edge device is also envisioned.

2.4 Blurb for public dissemination on UTTER's website

`pyannote.mobile` project led to the extension of the `pyannote.audio` open-source speaker diarization toolkit to perform speaker diarization in real-time while controlling the trade-off between latency and accuracy. It also led to the creation of an iOS/macOS streaming speaker diarization SDK which will be handed over to interested parties through the local university tech transfer office.

3 Recommendation by Project Sponsor

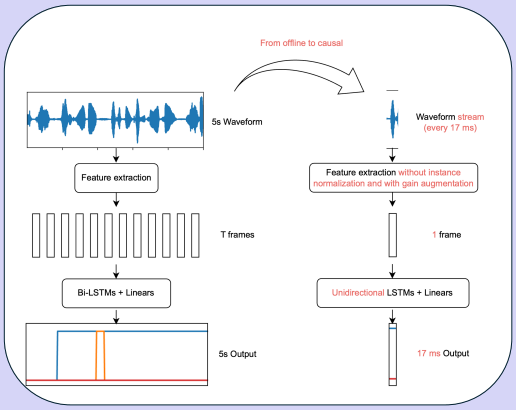
As the sponsor of this project, we confirm that the project successfully delivered its planned results. The dissemination efforts were effective, including an iOS application soon to be available, and a scientific paper published at Interspeech 2024. The project lead also provided comprehensive documentation and effective communication throughout the duration of the project, which were appropriate and well-aligned with the project's objectives. Overall, we recommend this project positively, as it has met its key objectives and demonstrated potential for future impact.

Multi-latency look-ahead for streaming speaker segmentation

Bilal Rahou Hervé Bredin

github.com/pyannote/pyannote-audio

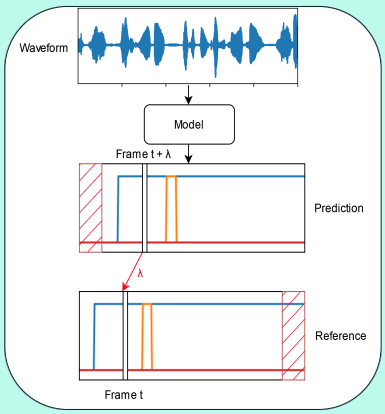
1 From offline to causal frame-wise segmentation model



The few changes (removing the instance normalization and the bidirectionality of the LSTMs) significantly worsen the performance of the model. The gain augmentation is added to compensate the lack of normalization. The table below summarizes the impact of these changes.

LSTM direction	Instance norm.	Gain augm.	5s chunk DER%
⟷	✓		17.3
⟶	✓		20.3
⟶			22.2
⟶		✓	21.1

2 Training with look-ahead



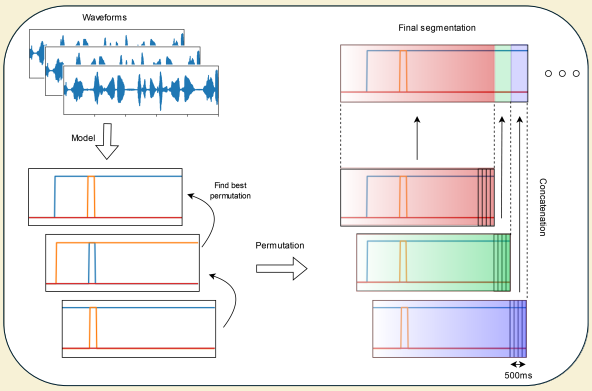
To improve the model's deteriorated quality, we introduce a look-ahead mechanism. The structure of the model does not change, the only change occurs during training, where the loss is calculated between shifted predictions and references. The shift λ corresponds to the added latency to the system. Below is the formula of the loss:

$$\mathcal{L}\left(\mathbf{y}, \hat{\mathbf{y}}\right) = \min_{p \in \mathbb{P}} \mathcal{L}_{\text{CE}}\left(p(\mathbf{y}_{0 \to T-\lambda}), \hat{\mathbf{y}}_{\lambda \to T}\right)$$

The approach can easily be generalized to multiple latencies, though that needs a slight modification of the final classifier layer. We duplicate the final classifier layer K times, so that the model now outputs K predictions, one for each latency. The rest of the model is shared by every latency. The training loss is computed as the sum of the aforementioned look-ahead training loss over each latency:

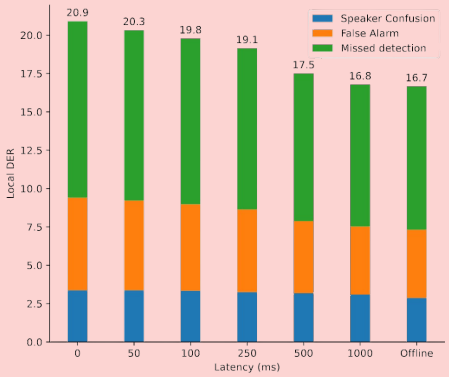
$$\mathcal{L}\left(\mathbf{y}, \hat{\mathbf{y}}\right) = \sum_{k=1}^{K} \min_{p \in \mathbb{P}} \mathcal{L}_{\text{CE}}\left(p(\mathbf{y}_{0 \to T-\lambda_k}), \hat{\mathbf{y}}^k_{\lambda_k \to T}\right)$$

3 Inference



To keep the advantages of hybrid speaker diarization approaches, we stick with an approach based on sliding windows (5s chunks with a 500ms stride). With the exception of the very first chunk that is used entirely, only the final 500ms of each subsequent chunk is used in the final concatenated output.

4 Results



As expected, the performance of the streaming system increases with the latency, almost closing the gap with its offline counterpart for a 1s latency (AMI).

Figure 2: Poster presented at Interspeech 2024

References

Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *INTERSPEECH 2023*, pages 1983–1987, 2023. doi: 10.21437/Interspeech.2023-105.

Hervé Bredin and Antoine Laurent. End-to-end speaker segmentation for overlap-aware resegmentation. In *Interspeech 2021*, pages 3111–3115, 2021. doi: 10.21437/Interspeech.2021-560.

Alexis Plaquet and Hervé Bredin. Powerset multi-class cross entropy loss for neural speaker diarization. In *INTERSPEECH 2023*, pages 3222–3226, 2023. doi: 10.21437/Interspeech.2023-205.

Bilal Rahou and Hervé Bredin. Multi-latency look-ahead for streaming speaker segmentation. In *Interspeech 2024*, pages 1610–1614, 2024. doi: 10.21437/Interspeech.2024-923.

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D6/D1.2 Report on first set of FSTP projects