Report on the outcomes of a Short-Term Scientific Mission[1]

Action number: CA18209

Grantee name: Gokhan Ozkan

Details of the STSM

Title: Analysis and Opening the Data of the Natural Language Processing Study for Worldwide Language Learning Apps and Platforms

Start and end date: 09/01/2023 to 23/01/2023

Description of the work carried out during the STSM

Description of the activities carried out during the STSM. Any deviations from the initial working plan shall also be described in this section.

(max. 500 words)

The project plan was to complete a final academic paper on natural language processing research on digital language learning solutions within two weeks. The first week focused on defining work packages, planning the distribution of tasks, identifying missing points in the draft article, and completing the literature review and state of the art sections. The second week focused on finalizing the article, describing the survey and research instruments in the methodology section, presenting the results of the data analysis in the results section, and discussing the contributions of the study and limitations in the discussion section. The paper was also prepared for submission to academic journals such as Computer Assisted Language Learning, International Journal of Computer Supported Collaborative Learning, Journal of the Learning Sciences, and Journal of Learning Analytics. Overall, the project aimed to provide valuable insights into the structure and properties of language data at a large scale and to contribute to the development of linked data technologies and natural language processing techniques for linguistic data science. Additionally, we have also conducted an extensive analysis on the current trends and best practices in digital language learning solutions, as well as identifying areas for future research. Furthermore, we also prepared the findings to present in a conference or a workshop and made recommendations for educators and practitioners to enhance the effectiveness of digital language learning.

Week 1:

[1] This report is submitted by the grantee to the Action MC for approval and for claiming payment of the awarded grant. The Grant Awarding Coordinator coordinates the evaluation of this report on behalf of the Action MC and instructs the GH for payment of the Grant.

Monday: Define work packages, plan distribution of tasks

Tuesday: Identify missing points in draft article, discuss target journal and obtain writing style guides

Wednesday: Complete comments and self-feedbacks in literature review and state of the art sections

Thursday: Prepare statistical tables for methodology and results section

Friday: Complete literature review and state-of-art section

Week 2:

Monday: Finalize article-in-progress with required writing style, work on identified missing points

Tuesday: Describe survey, questions, and research instruments in methodology section, describe data collection and analysis processes

Wednesday: Share core data with Dr. Lionel Nicolas for double-checking statistics and tables, collaborate on finalizing methodology and results sections

Thursday: Conclude with discussion of contributions to future work, limitations of the study, and overall discussion

Friday: Conduct academic ethical checks, finalize referencing and formatting, review and edit paper for target journal standards

Saturday: Prepare presentation for upcoming conference, finalize final version of paper for submission

Sunday: Practice conference presentation, complete any necessary administrative tasks.

Target journals selected for submission:

Computer Assisted Language Learning

International Journal of Computer Supported Collaborative Learning

The output: A finalized draft of academic paper on natural language processing research on digital language learning solutions that provides an overview of the analysis conducted on language learning apps and platforms, including a description of the datasets used, labelling and annotation process, and the most popular tools identified. The paper also included a methodology section, results section, and a discussion section on contributions to the field, limitations, and future work as the primary objective of this project was to prepare the study for submission to academic journals and make linguistic data accessible to field researchers.

Description of the STSM main achievements and planned follow-up activities

Description and assessment of whether the STSM achieved its planned goals and expected outcomes, including specific contribution to Action objective and deliverables, or publications resulting from the STSM. Agreed plans for future follow-up collaborations shall also be described in this section.

(max. 500 words)

Main achievements:

Completion of a finalized draft of academic paper on natural language processing research on digital language learning solutions, including an overview of the analysis conducted on language learning apps and platforms, a description of the datasets used, the labelling and annotation process, and the most popular tools identified. The paper includes a methodology section that describes the data collection and analysis processes in detail, as well as a results section that presents the findings of the study and a discussion section that discusses the contributions of the study to the field, as well as any limitations and areas for future work. The paper is of high enough quality to meet the standards of academic journals and will be submitted to targeted journals such as Computer Assisted Language Learning, ijCSCL, Journal of the Learning Sciences and Journal of Learning Analytics. The output of the research provided valuable insights into the structure and properties of language data at a large scale, which can be used to extract new knowledge and insights about linguistic data science and inform the development of a mature, holistic ecosystem of multilingual and semantically interoperable linguistic data.

Planned follow-up activity:

- Review and editing of the paper to ensure it meets the standards of the target journal.
- Creation of a presentation outlining the key points of the paper to be presented at an upcoming conference.
- Preparations for the conference presentation, including practicing the presentation and creating any necessary visual aids.
- Completing any necessary administrative tasks, such as completing reports or filling out paperwork related to the project.
- Submit the final version of the paper to the target journal for review.

Appendix 1. Parts Written During STSM

Phase II. Exploring Worldwide Platforms

Introduction

The site visit in Phase I revealed the requirement for a tool that may be utilized by all learners inside and outside the classroom. As the parts of this educational tool were being put together, another search for the parameters of similar tools on the market turned up, was planned and carried out. Business models, learner preferences, app types, generated materials' service platforms, and interactivity models were all considered essential parts of the process. As a result, the objective of this Phase was to define the preferences of the learners as well as the digital solutions that are now accessible on the market. In addition, the various types of apps had been through a process of rapid development, and the solution that was picked to be produced was the one that was the most adaptable and accessible. The researchers collaborated with a group of other researchers within the confines of a COST Action, assisted in the distribution and preparation of surveys, and analyzed the data collected along with data scientists and computer scientists in order to be able to provide a framework for the market by providing a snapshot of existing solutions and informing the environment in which a to-be-designed tool will be implemented.

Table. Phase II Research Questions and Process Details

DBR Step	Analysis
Research Question	How have the many functions and services that language learning applications and platforms provide, such as the languages that are taught, interactive exercises, user interaction, and business models, been utilized and put into practice in the current market?
Data Collection Instruments	Survey Annotation Form
Research Method	Survey
Intervention	No intervention, descriptive study
Target population	Digital language teaching app&platform users

Number of Participants	1374
Results obtained	A variety of provided languages, business model types, operating frameworks, and offerings of interactive exercises.

The table above describes a study that was conducted to examine the various features and services offered by language learning applications and platforms. The research question for the study was focused on how these features, including the languages taught, interactive exercises, and user interaction, are being used and implemented in the current market. To collect data for the study, we used a survey as the data collection instrument, and the target population was users of digital language teaching apps and platforms. The results of the study indicated that there is a diverse range of languages, business models, operating frameworks, and interactive exercises being offered by language learning platforms.

Research Goal

The goal of the Phase II study was to explore the populated language learning digital media solutions for language learning purposes and carry out frequency analysis in their size, platforms, business model, choice of interactivity in exercises and provided languages with the purpose of deciding upon the foundation principles. Information on entities provided by participants was annotated, and the existence of interactive exercises was noted for further analysis in future phases. Data regarding the size and business model of apps/platforms were aimed to indicate overall learner e-learning preferences. Information about the number of provided languages for teaching services was also collected in all apps/platforms with the purpose of detecting the dedicated products in the market. In addition, digital media platforms and app types were investigated with the purpose of comparing in terms of performance, flexibility, and universality.

Participants

The Phase II study enrolled 1374 people from 67 different countries with 42 different native languages. All participants confirmed they had used at least one digital language learning platform. Project team members contacted the participants by emailing a list of 182 committee members from 38 different countries in the project and asked for assistance with the survey distribution. Because the project's council was comprised of academics, the survey was disseminated mainly to university students.

Some demographics of participants were not considered significant, however, because the study's purpose was to identify users of digital language platforms.

Data Collection Instruments

This phase utilized two instruments. The first was the survey issued to participants, and the second was the annotation file for the responses collected.

In the survey distributed to members of the committee in the COST Action project, our working group firstly wrote a statement about the project and research, and participants were informed about the consent process and that, by completing the survey, they were assisting a research initiative and consenting to the use of their data for research purposes. The survey was dedicated to collecting the names and reference links of various digital language learning platforms. Participants were requested to provide up to ten URL links that they know and/or have encountered. Participants were also notified that they would be kept informed of the survey's results, the next steps of the research, project initiatives in general, as well as the survey's expiration date and project organizers' email addresses for any potential concerns.

The annotation file for the gathered responses served as the second instrument for conducting the data analysis. A number of meetings were held by the project's working group in order to select the probable dimensions that were going to be evaluated. Following that, the annotation file was consulted in order to ascertain the following variables:

- cryptic and working references,
- platform language,
- access blockers,
- size of the user base,
- existing interactive exercises,
- automatic feedback mechanism,
- number of languages taught,
- business model

This four-step process followed by the working group led to the investigation of the aforementioned dimensions for the existing global digital language learning platforms. The researchers took part in all stages of the project, including attending meetings to prepare the survey in Coimbra,

Portugal, as well as online meetings to prepare the annotation file. The researchers also annotated all items in the first round with a data scientist and then annotated all items in the second round with the team working group members. The researchers double-checked any changes between the first and second rounds at the University of Malta's AI department with a computer scientist from the working group.

Procedure

The data collection instruments were prepared by researchers in a meeting held at the University of Coimbra, Portugal, within the context of the COST Action-European Network for Combining Language Learning with Crowdsourcing Techniques (Action Nu.: CA16105). The technological requirements for survey distribution were relatively basic. Therefore, most survey technologies should be able to accomplish the work. Despite this, the Cost Action working group considered and tested various options before settling on the lime survey as the most efficient one for our case.

To motivate participants, we held a lottery (e.g., to win Amazon vouchers). A "ticket" was awarded to each participant for their response. The value of these tickets increased if the reference to the language-learning platform was uncommon. Users were encouraged to submit references to smaller or more country-specific platforms as a result of this change. We limited the number of participant contributions to 10 answers to generate a sense of doability and forced them to refrain from spamming us with well-known platforms.

We asked for only the most essential information and concentrated almost entirely on acquiring their URLs. If participants utilize some platforms, we asked their view of the platform's utility and the frequency with which they use it. Since this survey was the most extensively disseminated component of this effort, we wanted to ensure its accuracy before sending it out. Prior to the distribution of the survey, we thus conducted a pilot study and examined the potential communication channels as below:

- Cost Action member of committee (MC) group and social media,
- Teachers' associations,
- Universities students,
- Citizen Science platforms,
- Search engine scanning,

- Local conferences.

Hyperlinks of the surveys were distributed online to 1702 participants in 42 different languages from 67 countries, and items of 576 different digital solutions were collected and listed. Following this process, the working group in Cost Action decided that the researchers would annotate 595 responses for the detection of the populated digital language-learning platforms. Annotation of items planned to be organized in four stages. The first stage involved the creation of an annotation file that accounted for the analysis's dimensions. The second stage included the first annotation of all items at once by the researchers. The third phase was determined to be the re-annotation of all items by all group members in order to check the first annotation phase, increase consistency, and inform the researchers of any discrepancies so that he can check and change the different annotations compared to his in the first stage. In the final stage, referred to as stage four, the researchers concentrated on the inconsistencies, provided justifications for the final decisions, and obtained confirmation of this analysis from the working group leader. For the early annotation, the researchers worked with a data scientist at the Bulgarian Academy of Sciences to analyze each item on the annotation form. For the reannotation stage, the researchers collaborated with a computer scientist at the University of Malta's Artificial Intelligence Department to ensure analysis consistency and reliability. COST Action-CA16105 funded all the costs for the process within the scope of short-term scientific mission (STSM) projects.

Responses From Participants

The study process consists of two parts, the first of which was to compile a list of references to language-learning platforms and the second of which was to tag and analyze each one in-depth with an annotation form. The numbers reached during the study were determined to be as follows after the valid references were extracted from the total data and the cryptic links were eliminated. The number of total references was 595, but 31 responses out of them were deemed invalid and were taken out of the study.

Table. Number of Valid Responses Without Blockers

	Number of Blockers	**Total**
Number of references		595
Invalid references	31	564
Language barrier	162	402

Registration and other blockers	9	393

The analysis was subsequently revised to remove platform links to languages that were unknown to the researchers (n = 162). The number of references made in the Dutch language was the highest among links containing an unknown language. s came across a total of 25 distinct languages that were completely beyond his comprehension when he was looking through the reference links. It is essential to highlight the fact that the use of translation technologies was found to be ineffective in terms of comprehending the target language. As a direct consequence, the researchers and the working group concluded that they should be excluded from the research. This decision was made since the analysis was meant to be as accurate as reasonably practicable.

Table. Reference Links in an Unintelligible Language to the Annotator

Language	Number of Entities (n)
Croatian	15
Dutch	13
Portuguese	11
French	9
Hebrew	9
Spanish (Castilian)	9
Basque	8
Other languages	88
Total	162

The requirement to register with a credit card, and make a monthly or annual payment to gain access to the website and understand the characteristic of the learning offer (n = 9) were additional obstacles that the researchers encountered. The researchers were only able to gain access to some of the platforms that need payment; however, not all of them. The working group came to the conclusion that blockers with references should not be annotated as a result of these challenges.

Table. Number of Blockers Among the Reference Links

Blockers	Number (n)

Credit card registration	9
Invalid	393
Total	402

The requirement to register a credit card for recurring monthly or annual payments was not the only obstacle in the way of gaining access to the platform. There were also other reasons, such as broken links within the platform and the need for access codes to MOOCs. As a result of the dynamic nature of the digital language learning platforms and the fact that the market necessitates substantial updates in order for these platforms to catch up with the modern interfaces, methods, and standards of service, the researchers came across a few websites that had broken links in various points of entry to the website.

Table. List of Issues Preventing the Access to the Platform

Reason	**Number (n)**
Available Soon	1
Broken sublinks	1
Class/Book Code	1
Closed Server/No more working	1
Registration Code	1
Total	5

The researchers also examined the user bases of platforms to identify the most populated ones available on the market and the distribution among three coarse categories: small (less than 10 000 users), medium (between 10 000 and 1 000 000 users), and large (more than 1 000 000 users). Excluding the option "I don't know/I can't tell." (31.31%), 9.6% of all platforms were small, 31.82 % were medium-sized, and 27.27 % were large. The decision regarding the size of the user base was taken through a comprehensive review of the platforms, particularly their social media followers, likes and download statistics.

Table. Main characteristics of non-language-learning entities

Tags	**Number (n)**	**Percentage (%)**

Website to create interactive exercises.	1	0.83%
Chatbot	1	0.83%
Dictionary	14	11.57%
Find a tutor	1	0.83%
HTML language	1	0.83%
Irrelevant	1	0.83%
Learning Management System	20	16.53%
Language Learning School	1	0.83%
Music (or Podcast) platform	1	0.83%
Notetaking/Flashcard App	1	0.83%
Reading Assistant	15	12.40%
Social Media/Language Exchange	1	0.83%
Webinar Platform	63	52.07%
Website to share or download language learning content	1	0.83%

Some of the reference links go to irrelevant websites, websites of language learning institutions, content sharing types of multimedia platforms, etc., and not all platforms were giving interactive exercises with automatic feedback mechanisms. However, participants opted to provide a link to any and all digital solutions that aid them in learning the target language. Tools for webinars and dictionaries could fall within this category.

Due to the fact that the chain questions were utilized for the annotation form, as well as annotation disagreements that arose from a question to the other, the number of responses to each question may vary. Therefore, not all questions were displayed on each platform. For instance, questions concerning business models or service platforms were not displayed on platforms with cryptic references or irrelevant links.

Table 25. Operating Framework of the Entities

Service Platform	**Number (n)**	**Percentage (%)**

Online website	175	43.64
Android app	109	27.18
IOS app	96	23.94
Windows app	7	1.75
Mac app	11	2.74
Other	3	0.75
Total	401	100.00

In order to frame another aspect that was found significant, it was necessary to comprehend the business models of the entities. This dimension was included for no other reason than to have a complete picture of the market as a whole. Within the annotation item pertaining to the business model, there were three different options available. "Mainly free" refers to either entirely free or ad-included models, whilst "mostly not free" versions usually only provide a limited demo and require payment before you can advance with any further activities. Lastly, the business model marked as "partially free" does permit you to use the platform without paying and restricts your usage at some point.

Table. Business Models of Language Learning Platforms with Automatic Feedback Mechanisms

Business Models	**Number (n)**	**Percentage (%)**
Mostly free (e.g., the vast majority of content and features are freely accessible)	100	50.00%
Mostly NOT free (e.g., you can have a demo or a tour of the tools and content, but most of them are not available unless you pay for it)	48	24.00%

Partially free (e.g., paying some money gives you access to additional content or features or allows you to use the service more often)	52	26.00%
Total	200	

This question could only be shown on 200 entities because that is the number of identified entities (entities identified as language-learning platforms with automative feedback mechanisms and no blockers) that have been consensually annotated twice. It could be concluded from the results that the market favors free services supported by adverts and/or other income sources unless entirely free (50.00%).

The final aspect of the market for digital language-learning solutions examined was the frequency of languages taught among the reference links collected by the participants. A total of 312 languages were encountered through the use of digital language learning platforms. This data represents the number of languages that are taught by digital language learning programs or supported by these platforms. Below are the most common languages found in reference links.

Table 27. Popular Languages in Digital Language Learning Solutions Market

Languages	Number (n)
English	135
German	85
Spanish	73
Portuguese	67
French	66
Italian	65
Russian	56
Japanese	55
Dutch	51
Swedish	50

The general framework of digital language learning solutions revealed that digitalized language learning solutions either offer English as a language service or already teach it. The most popular service platforms are online websites and mobile applications; more than half of them use advertising on free versions as their business model. However, desktop versions and alternative services are used so infrequently, and paid versions of digital services are not widely employed. Consequently, each entity was analyzed in depth, and design decisions for the instructional tool were planned to be developed with a great deal of consideration for existing digital language learning solutions.

Conclusions

This phase's objective was to define the existing digital language learning solutions on the market in order to comprehend the services provided and learner preferences. The inclusion of this stage in the research project was done with the intention of assisting the researchers in better comprehending the implications and uses of digital solutions. Valid references, languages taught, business models, service platforms, user size, and interactive examples with an automatic feedback mechanism make up the aspects of analysis. Based on the responses and the statistics that were collected, it was determined that English is the most used language, with German and Spanish coming in second and third, respectively. In terms of business models, freemium solutions were the most popular. There is a wide variety of service platforms available, but mobile compatibility is vital due to the widespread use of platforms like iOS and Android. In the final step of this research project, user sizes and interactive examples are studied for a follow-up study to better comprehend the proportions of exercises found in textbooks and digital solutions.

Additionally, the ideal service platform for the educational tool that was developed for this study was researched. iOS, Android, and Windows versions were all favored by different learners, indicating that both the learners' preferences and their devices are diverse. As a result, the universality and adaptability of the design were considered to be of the utmost importance in order to reach all students within the same educational setting.

As an implication of the findings from this stage of the research, it was discovered that a website or access to a learning management system (LMS) was a practical means to reach all learners with a smart device. This was owing to the fact that all smart devices should have an active

internet connection that can access websites, and all students in the educational setting have access to this technology. The solution, therefore, met the universality criterion but was not flexible enough to accommodate the growing popularity of mobile-friendly digital language learning platforms over traditional websites or learning management systems. Because of this, the researchers opted to test both native and hybrid app solutions for mobile versions of the application in order to satisfy the adaptability criterion. Many content management systems, or CMSs, were researched to achieve this. The researchers decided to proceed with the most reliable one. Following this, research was conducted on learning management systems (LMS), and the most versatile and stable system with device compatibility was chosen to embed into CMS. The LMS-embedded CMS concept was initially implemented as a web view app so that its effectiveness could be compared to that of hybrid versions. The researchers then encountered two disadvantages of this method: incompatibility with software updates and ii) a slow operating system. As an alternative, the hybrid solution developed during the conceptual design phase was ultimately implemented in the form of a Progressive Web App (PWA). The PWA solution provided the most advantages in terms of universality, adaptability, flexibility, speed, and software update.

Appendix 2. Summary by Lionel Nicolas

Towards Mapping the Landscape of Existing Language Learning Solutions Offering Interactive Exercises

Lionel Nicolas, Institute Applied for Linguistics, Eurac Research, Bolzano, Italy
Gokhan Ozkan, School of Foreign Languages, Kırklareli University, Turkey

This work presents the ongoing efforts and current results achieved in identifying and analysing language learning solutions that offer interactive exercises through a large survey run in the context and with the support of the enetCollect COST Action (European Network for Combining Language Learning with Crowdsourcing Techniques). The survey had 638 participants who provided 595 different references, through which we identified over 250 language learning platforms offering interactive exercises.

Such mapping efforts are originally aimed at supporting the exploration of an interdisciplinary approach combining Crowdsourcing, Computer Assisted Language Learning (CALL) and Natural Language Processing (NLP), which tackles two major CALL and NLP issues: the lack of exercise content for CALL on the one hand and the lack of language-related datasets in NLP on the other hand. This approach relies on the generic idea of combining a specific type of exercise (e.g. a vocabulary exercise) with a language-related dataset (e.g. an NLP lexicon) from which the content for this type of exercise can be generated automatically. The answers to the exercises generated are then crowdsourced and used to improve the dataset (e.g. validate/discard/add an entry to the NLP lexicon). It exploits the fact that, on a conceptual level, a learner studying a language and a stakeholder curating a language-related dataset (e.g. an NLP researcher) are actually doing two similar tasks as they are both curating a language model. A decisive factor for implementing such a crowdsourcing approach is to implement it into an existing user workflow such as an existing language learning solution, which is the reason why this initiative was originally undertaken.

Despite the fact that our efforts were originally targeted to the aforementioned purpose, they also provided us the opportunity to study the language learning platforms available nowadays with a coverage that is unmatched in the state of the art we identified, such as Heil *et al.* (2016) and Kukulska‑Hulme & Viberg (2018). We thus not only identified language learning solutions but also registered several characteristics such as, among others, languages covered, popularity, business model, type of interfaces (e.g. website, android app, etc.).

In our presentation, we will briefly present the overall initiative that has motivated our efforts, discuss how we devised the survey, its distribution, the characteristics of the set of answers we obtained and how we post-annotated the references. Finally, we will provide the current conclusions we derived from the statistics we computed on the annotations including, among other things, our current conclusions regarding the number of languages covered by the solutions, the languages best covered by them and the correlation between the size of the speaker communities of the languages and the number of solutions.