IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 September 2023, accepted 8 October 2023, date of publication 13 October 2023, date of current version 19 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3324390

RESEARCH ARTICLE

Visualizing the Behavior of Learning European Portuguese in Different Regions of the World Through a Mobile Application

LAP MAN HOI[1], **(Member, IEEE), YUQI SUN**[1], **WEI KE**[1], **(Member, IEEE), AND SIO KEI IM**[1,2], **(Member, IEEE)**

[1]Faculty of Applied Sciences, Macao Polytechnic University, Macau, China
[2]Engineering Research Centre of Applied Technology on Machine Translation and Artificial Intelligence, Ministry of Education, Macao Polytechnic University, Macau, China

Corresponding author: Lap Man Hoi (lmhoi@mpu.edu.mo)

This work was supported by Macao Polytechnic University under Project RP/ESCA-03/2020.

ABSTRACT This study explores the "Diz lá!" mobile application, an innovative tool released in 2018 that aims to enable users, especially Chinese speakers, to learn Portuguese. This mobile application harnesses the principles of *Mobile-Assisted Language Learning* (MALL) and *Self-Determination Theory* (SDT), facilitating continued language education amid the COVID-19 pandemic. Our research exploits user habits reflected in more than five years of data to build multi-dimensional models for visualizing large datasets, with a focus on learning patterns related to verb conjugation. Insights reveal that most users are language school students from Macao, China, and Portuguese-speaking counties/regions, with a remarkable preference for learning Portuguese verb conjugations. The research results also show that people like to learn verbs for comparison. Notably, despite the pandemic, an upward trend was observed in the learning of these conjugations. Our findings offer crucial implications for the design of pedagogical strategies and refinement of language learning (MALL) apps, highlighting areas of difficulty and learner preferences. We also used *machine learning* (ML) technologies to create a predictive model to recommend relevant learning materials to users. As a result, this study stands at the intersection of technology-enhanced language learning and educational research, demonstrating how they can synergistically contribute to optimizing language learning outcomes.

INDEX TERMS Conjugator, COVID-19 pandemic, data visualization, Diz lá!, European Portuguese, learning patterns, mobile-assisted language learning, second language acquisition, self-determination theory, star schema data model.

I. INTRODUCTION

As we progress further into the digital age, the integration of technology in education has become increasingly significant, with mobile applications (apps) playing a pivotal role in the learning process. The onset of the *Coronavirus Disease 2019* (COVID-19) pandemic [1] catalyzed the digital shift, and MALL has gained particular momentum in the post-pandemic era.

In Macao, a *Special Administrative Region* (SAR) of the People's Republic of China, where Chinese and Portuguese are the official languages, the need for quality language learning resources is palpable. Universities in Macao, such as *Macao Polytechnic University* (MPU), are renowned in Asia for their Chinese and Portuguese studies. With the continuously increasing number of *Massive Open Online Courses* (MOOCs) on the Internet, various electronic learning materials that are convenient for students and teachers have been developed to enhance the educational experience.

The associate editor coordinating the review of this manuscript and approving it for publication was S. Chandrasekaran.

In the post-COVID era, the rapid migration of digital platforms has foregrounded the importance of MALL. These tools amalgamate the principles of traditional pedagogy with the adaptability and accessibility of modern technology. Amid this context, the "Diz lá!" app emerges as a beacon of innovation. This is an example of how academic research can be effectively transformed into results, enhancing the ability to innovate and apply state-of-the-art technologies.

In light of the COVID-19 pandemic, which has put many traditional in-person classes on hold, the capacity for student autonomy in learning has been elevated. Distance and online learning solutions, such as language learning apps, are becoming increasingly important. Here, this mobile app adopts MALL principles and is guided by SDT to promote this autonomy by enabling continuous language education within the constraints caused by the pandemic.

In this article, we demonstrate how to construct a data pond that integrates all possible data sources. This data pond is the foundation of our research environment, which provides a data pipeline to build star data models for data visualization and output text files for training artificial intelligence models. We specialize in visualizing and analyzing large amounts of user data accumulated from a mobile app. By utilizing modern visualization techniques, we hope to shed light on user learning behavior, specifically verb conjugation. This approach, rooted in big data concepts, aims to allow us to better understand these behaviors and then feed this back into the improvement of language learning resources. Moreover, additional functionality using machine learning technology to suggest relevant verbs for users to learn is also proposed and demonstrated. The motivation of this study is to use research data to explore the difficulties faced by the general public when learning Portuguese, use visual methods to present some learning patterns and use artificial intelligence technology to recommend relevant learning materials. Hopefully, these studies will be helpful to other educational mobile software developers.

The article is structured as follows: Section I briefly introduces the purpose of this study; Section II reviews state-of-the-art techniques and related research; Section III shows how to produce visual charts and tables, and how to build predictive AI models; Section IV presents all preliminary findings; Section V discusses and analyzes the significance of all results; Section VI is the conclusion and future work.

II. LITERATURE REVIEW

A. MOBILE ASSISTED LANGUAGE LEARNING

The development of mobile technology has contributed to a paradigm shift in language learning methods. *Mobile Assisted Language Learning* (MALL), defined by Kukullska-Hulme [2] as the use of smartphones and other portable devices in language learning environments, has attracted considerable attention, particularly because of its potential to provide unique, high-quality educational opportunities [3].

Research emphasizes the importance of understanding how students employ these technologies to supplement language learning and highlights the transformative role of MALL [4]. The benefits that mobile technology brings to education are manifold, such as instant access to learning materials, portability, and facilitating a personalized learning experience. These technologies also offer a broader range of opportunities for foreign language learning, and their inherent characteristics of flexibility, affordability, and user-friendliness make them attractive learning tools [5], [6], [7].

Furthermore, MALL enables learners to control their own learning process, thereby increasing their language learning motivation, autonomy, and self-direction [8]. By allowing learners to interact with content that matches their level of competency and preferred style, MALL fosters a personalized learning experience that surpasses what traditional classroom settings can offer [9]. This paradigm shift further contributes to a transition from teacher-centric to learner-centric approaches that empower learners to control their learning pace and environment.

Interestingly, the prevalence of action learning is most prominent in higher education settings, with more than half of all action learning research conducted in this context [10]. A series of studies devoted to second and foreign-language mobile learning have shown that mobile technologies can effectively support various aspects of language learning [11], [12], [13]. However, despite these advances, the exploration of how best to leverage mobile technologies for language learning continues, highlighting the need for continued research in this area [14].

The recent COVID-19 pandemic has further highlighted the importance of MALL in tertiary education. Most studies conducted during this period focused on analyzing the opinions and experiences of participants [15]. Despite the generally positive reception to mobile technologies for online language learning [16], research has often neglected the impact of different mobile technologies and user contexts on these outcomes [17]. In the case of China, WeChat, a platform popular among university students, is particularly effective in supporting English learning [18]. However, the challenges of vocabulary acquisition among Chinese English learners, especially during the pandemic, remain under-investigated [19]. This observation underscores the need for a nuanced understanding of MALL, taking into account the diversity of applications, user contexts, and learning outcomes across academic settings and time, especially before, during, and after global crises such as the COVID-19 pandemic.

B. SELF-DETERMINATION THEORY

Self-Determination Theory (SDT), formulated by Deci and Ryan [20], [21], provides an insightful lens through which we can understand learner engagement in MALL. SDT prioritizes three basic psychological needs—autonomy, competence, and relatedness—as key drivers of motivation, and their fulfillment is essential for successful language learning [20], [21], [22], [23].

Within this triad, autonomy - viewed as the learner's ability to control their own learning journey - occupies a pivotal place [21], [23]. Autonomy in language learning is described as the ability of learners to control their own learning [24]. In a traditional classroom setting, this control often comes with constraints. However, the proliferation of MALL applications has reshaped the learning environment, providing learners with greater flexibility and control. For instance, learners can decide what, when, and how to learn, thereby enhancing their sense of autonomy. This amplified sense of autonomy significantly enhances learner engagement and promotes more successful language acquisition [2], [25].

The COVID-19 pandemic has highlighted the need for self-directed learning tools as global restrictions have led to an urgent shift to online learning, boosting the adoption of MALL applications [26], [27]. Not only can these apps replace in-person instruction, but they can also effectively address the growing need for autonomy-enhancing tools during a time of unprecedented educational disruption [28]. Given its emphasis on autonomy, SDT provides a valuable lens through which to understand how MALL can support self-driven, participatory learning during such challenging times. Consequently, in the face of the drastic changes caused by the epidemic, MALL has become an important strategy to meet learners' autonomous needs in the language learning process.

Moreover, given the constraints imposed by the pandemic, autonomous learning through MALL can also help fulfill the needs for competence (sense of efficacy) and relatedness (sense of caring relationships), two other critical components of SDT [21], [22]. MALL applications that incorporate interactive features or community elements can help learners feel connected and competent in their language learning journey, thereby promoting intrinsic motivation [6], [23].

C. TIME SERIES DATABASE

A *Time Series Database* (TSDB) is designed to store and retrieve data records that are part of a "time series" which is a set of data points associated with a timestamp [29]. Typical TSDB-type database engines are InfluxDB and TimescaleDB.

With the rapid growth of *Internet of Things* (IoT) projects, the demand for TSDB is also increasing. TSDB is used to handle transactions that involve intensive, large-scale insertion of data records. For example, IoT activity logs track the status of sensors every second. These frequent data inserts will create many append table paging operations and may cause the database engine to crash. Uber's case study [30] shows that traditional *Relational Database Management System* (RDBMS) are no longer able to cope with IoT projects that require inserting large amounts of data into database tables.

For that reason, TimescaleDB uses Hypertable and Chunks technology to enhance performance. Data insertion can be accomplished by parallelizing chunks across clusters or disks based on specific partition keys, and complex queries can be optimized by leveraging metadata for each chunk [29].

The data of this mobile app are mainly stored in log files. Since they are not in a well-structured format, TSDB-type database engines are best suited to manipulate such data items.

D. DATA VISUALIZATION

As the concept of big data becomes more practical, people use data lakes and data warehousing to manipulate their important assets "data". However, one might argue that we should focus on visualizing data for analysis rather than just storing it. Data visualization is a form of visual communication that presents data in graphical form [31]. It helps provide a holistic view of large amounts of data to discover hidden content [32]. People cannot discover patterns through traditional tabular format data. Therefore, visualization diagrams in various formats are needed. Commonly used visual diagrams are area, bar, boxplot, bubble, dependency wheel, gauges, heatmap, line, network, pie, radar, scatter, sunburst, and treegraph charts.

However, traditional charts cannot fully meet the requirements of constructing complex, multi-dimensional, and large-scale data charts. Therefore, the data visualization process requires more advanced technology, which may involve the field of *computer graphics* (CG). Famous CG technologies include *Open Computing Language* (OpenCL), *Open Multi-Platform* (OpenMP), *Open Accelerator* (OpenACC), *Open Graphic Library* (OpenGL), WebGL, Direct3D, etc. [33], [34]. The specific usage depends on the hardware and software environment. CG developers often program in C, but this traditional programming language is complex and difficult to master.

Fortunately, there are many tools for rendering advanced charts (SAS, SOFA, R, Minitab, D3, Python, Javascript, etc.), and some research suggests presentation methods that better render charts for human readability (font size, color, size, layout, etc.) [34], [35]. Recently, people have been using data visualization techniques to build dashboards to monitor sales performance and thereby forecast and anticipate market trends [36]. Therefore, we strongly believe that hidden information can be discovered by visualizing massive amounts of data.

E. ANTICIPATING WORDS THROUGH ML

The verb conjugation is one of the difficulties that most students face when learning Portuguese. Similar conjugation rules and verbs (spelling, pronunciation, meaning, etc.) can be confusing. It is a common practice for people to put similar verbs together for comparative study. Therefore, we wanted to provide the app with a new feature that intelligently recommends some relevant verbs to learn.

Over the years, different approaches to machine-suggesting words have emerged. In the field of *Natural Language Processing* (NLP), "Next Word Prediction" has always been one of the important research areas for machines to imitate human speech.

There are different algorithms (bag of words, n-gram, word2vec, etc.) to handle this problem [37]. Relying on modern deep learning technology [38], the capabilities of *Long Short-Term Memory* (LSTM) and *Gated Recurrent Unit* (GRU) networks never cease to surprise us. These types of predictive models are constructed by training on large amounts of sentence-based data in an unsupervised manner. The machine learns the context of a sentence to anticipate the next word that may appear sequentially in the sentence.

Large language models (LLMs), which have received much attention recently, can even amaze us. It has had a huge impact in the field of NLP. Its ability to imitate human speech is sometimes indistinguishable. Some notable LLMs include OpenAI's *Generative Pre-trained Transformer* (GPT) [39], Google's *Bidirectional Encoder Representations from Transformers* (BERT) [40], and *Language Model for Dialogue Applications* (LaMDA) [41]. These LLMs have become one of the hottest research topics at present.

III. METHODOLOGY

This mobile app was developed by a collaborative effort of language and computer experts [42]. The present version of the app is designed to foster language learning in the digital age, providing learners with resources such as "Daily Word", "Conversation", "Vocabulary", "Video", and particularly the "Verb" section which is the most pertinent for our current discussion. This section boasts a vast repository of over 15,000 Portuguese verbs along with more than 100 verb conjugation rules [43]. To enhance the learner experience, it also includes human-recorded pronunciations of over 400 verbs by native European Portuguese speakers, significantly simplifying the understanding of complex and often tricky verb conjugations.

User anonymity is preserved and the information collected is used for educational purposes only. The data collected complies with big data principles and includes various sources such as system log files, transaction database records, app reviews in app stores, etc. The sheer volume and complexity of this data may seem daunting. However, through the application of data mining and visualization technologies, we can uncover meaningful patterns and narratives hidden beneath the surface [44].

Our methodology in this study concentrated on the visualization and in-depth analysis of this substantial dataset. We have been collecting data for more than five years, and now it's time for the data to tell us something. In order to produce the visual charts and tables required for the analysis, we performed a series of data processing steps.

Fig. 1 illustrates the workflow of the entire process. Users generate transaction data through mobile applications. Data from log files and database records are *extracted, transformed, and loaded* (ETL) into a central database or "data pond". The data were then fit into a multi-dimensional star schema material model for visualization and analysis. Visual diagrams are generated using JavaScript libraries. At the same time, modern machine learning technologies are used to build prediction models based on training and testing data in the data pond.

Given the emphasis of this study on data visualization and analysis, the following section summarizes the core procedures we employed and omits some minor details.

A. DATA PRE-PROCESSING

1) DATA COLLECTION

If we want to integrate all the data together according to the concept of big data, it is necessary to identify all possible sources of data in order to produce a holistic view for in-depth analysis. Some useful sources of data include mobile application online store monthly reports, web application server (Apache Tomcat) access log files, and *Online Transaction Processing* (OLTP) type database records.

- Online store reports provide information on the number of users, software crash records, device models, and app installs and uninstalls.
- Apache Tomcat log files provide user access information, including time, *Internet Protocol* (IP) address, device model, operating system version, web browser engine version, etc.
- Database records provide more detailed information about which function was accessed at what time.

Among them, log files are the most useful and comprehensive for in-depth analysis. Web application servers are configured by default to log limited user access information. There are a number of parameters, such as the "pattern" attribute [45], that can be configured to display user access details.

2) ADDITIONAL INFORMATION

Geolocation data can tell the user's location (country, city, latitude, longitude, etc.). This information is important for classifying user behavior by region. Geolocation data is obtained by parsing IP addresses. This parsing process requires an IP address resolver. There are many open-source APIs on the web (iplocation.net, app.ipgeolocation.io, api.ipbase.com, ip-api.com, etc.), and we chose ip-api.com [46] for our research.

Generally speaking, an IP address can reveal the location, but it may not truly accurately represent the user. Some scenarios are people traveling to other countries, or people using *Virtual Private Networks* (VPNs) to access applications through virtual IPs.

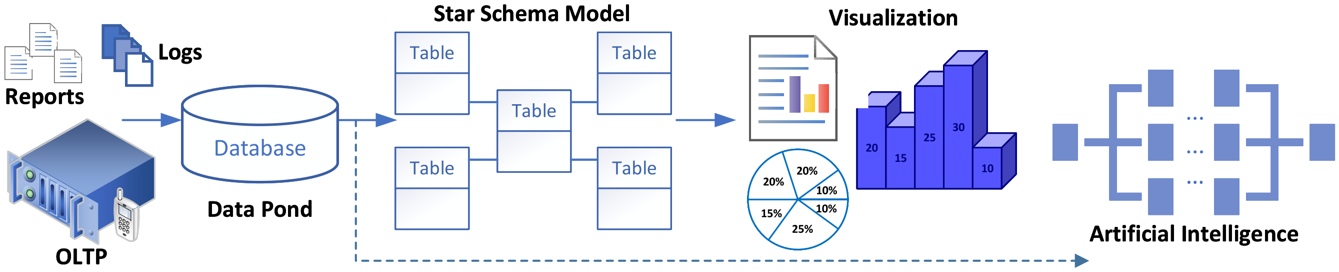


FIGURE 1. The workflow of building visualization and artificial intelligence environment.

Other scenarios are web robots scraping the content of mobile apps. If an IP address accesses the application very frequently in a short period of time, we can identify it as a robot and will not count the records generated by this IP address. Googlebot is Google's web crawling robot, and crawling is the process by which Googlebot discovers new and updated pages to add to the Google index [47]. It is not hard to see that Google is crawling our app on a regular basis. Consequently, the crawled content can be searched by the Google search engine, making the application known to the world. Geolocation data not only provides additional information about the user but also helps ignore the visit history of web spiders.

3) EXTRACT-TRANSFORM-LOAD

Since the data in our data sources are in different formats (spreadsheet-type files, text files, database records, etc.), cumbersome and traditional ETL processes are inevitable. Most of the data comes from Apache Tomcat log files, and the "pattern" attribute defines the structure of the log. Therefore, we can parse the information according to the official Apache manuals [45]. With Python's powerful data processing capabilities, we can easily ETL data into the data pond.

B. DATA POND

Gorelik clarified the differences between data puddles, data ponds, data lakes, and data oceans based on their size, usage, and maturity [48]. The term "data pond" is the most suitable to apply in our study. During the pilot study phase, we used the MySQL database engine to create a central database as a data pond to store all data in a database format. MySQL is one of the famous RDBMS-type database engines, now owned by Oracle Corporation. It is powerful, efficient, cross-platform, and lightweight for handling OLTP [49]. However, our data was collected from February 2008 to June 2023, more specifically, exactly 1,964 days and nearly 12.5 million records. The reading and writing process of MySQL is very time-consuming. If the database table index cannot be used, some statistical calculations may even take an hour to complete. Therefore, we switched to the TSDB engine to handle large amounts of data. Among all the TSDBs on the market, we chose TimescaleDB [50] because it is open-source and easy to use.

Therefore, data from all different sources are converted into database records, and each record is stored with a timestamp. The purpose of constructing a data pond is to keep all source data in an organized manner, which can be converted into a tight star structure for visualization, or output to a loose file structure to build an AI model. The logical diagram of our experimental environment is shown in Fig. 1.

C. MULTI-DIMENSIONAL DATA MODEL

1) STAR SCHEMA

After the data pond is built, we design a star schema data model and load the data into different dimensions and fact tables. Star schema is a collection of dimension tables and fact tables used for data analysis [51]. As a result, a time series repository is created that contains four-dimensional tables (time, learning material, geographical location, and device) and a fact table to hold the numeric values.

After studying the data pond, we are able to retrieve data on different types of content, mainly including the following dimensions:

- **Access time:** the timestamp (year, month, day, hour, minute, second, weekday, holiday, etc.) accessed by users.
- **Learning material:** the pages or functions (conversation, daily word, verb, video, and vocabulary) of the app accessed by users.
- **Geographic location:** the IP address used by the users. The IP address contains the user's geographical location information (country, region, latitude, longitude, etc.).
- **Device:** The devices (phone, tablet, TV, etc.) and platforms (Android, iOS, etc.) used by the users.

As a result, the star schema data model consists of the upper four dimensions for in-depth analysis.

2) USER SESSION

In order for this mobile application to retrieve learning data from a server-side OLTP database, we implemented a RESTful web service. Java Jersey is an open RESTful framework for developing stateless web services to operate databases. In REST architecture, the server does not retain the client state. This restriction is called statelessness [52].

Therefore, this mobile app is designed to allow users to access it without an account.

However, without user accounts, it would be difficult to track user activity (or sessions) over time. Because IP addresses log every action a user takes, we can use the time intervals between IP addresses to determine user sessions. Different server engines (web servers, application servers, database servers, etc.) have different default timeout values for user sessions, ranging from minutes to hours. Consequently, we define the user session interval as 30 minutes.

On the other hand, if two campaigns from the same IP last longer than 30 minutes, we will treat them as two different visits. If the time between each activity is less than 30 minutes, and there are many consecutive activities, the user session may be very long. Subsequently, we can then perform some interesting statistics on user sessions. For example, the relevance of the learning material during the session.

- **Learning Group**

If a user learns certain material during a session (such as verb conjugations), we can group this material into learning groups. Taking verbs as an example, learning groups only contain different verbs, even if a particular verb is learned multiple times in one session. Moreover, a group must contain more than one verb. Equation (1) denotes a group consisting of some verbs.

$$g(n) = \{v_1, v_2, \ldots v_n | v \in V\} \tag{1}$$

where g is a group, n is the group number, v_n is a distinct verb, and V is the set of all verbs.

- **Statistics on Learning Group**

After defining the learning groups, we want to know the frequency of the groups. This mobile app can help us understand the difficulties of learning Portuguese around the world if we can discover groups of people in different areas who are learning. Equation (2) denotes the frequency of a certain group at a specific time and region.

$$f(g, t, r) = x \tag{2}$$

where f is the frequency of the group g, t is a time period, r is a region, and x is the frequency value in integer.

Since verbs can exist in different groups, we can visualize the relationship of verbs to different groups. Equation (3) denotes the correlation of verbs with other groups.

$$Corr.(v) = \{g | g \in G, v \in g, g \neq \{v\}\} \tag{3}$$

where G is the set of all possible groups, v is a verb belonging to group g, and g cannot consist of only one verb.

Moreover, we can use (4) to find the most commonly learned groups across time and regions.

$$\underset{g}{\arg\max} \left\| \{g \in G | f(g, t, r) = x\} \right\| \tag{4}$$

The above-mentioned mechanism for determining user sessions is mainly based on time and IP address. It might not be very accurate, but it covers all data without forcing users to create their own accounts to access mobile apps.

D. VISUALIZING THE STAR SCHEMA

The beauty of a star schema is that it provides an intuitive answer to our data analysts' questions. Through this multi-dimensional data model, people can easily combine four dimensions (access time, learning materials, geographical location, and device) for queries. Data analysts can even twist dimensions and drill up and down to ask insight questions to uncover valuable information.

- Which mobile phone platform (**Device**) has the most access in China (**GeoLocation**)?
- Compare the number of hits on learning Portuguese verbs (**Learning Materials**) in 2018 and 2019 (**Time**).
- Which Portuguese verbs (**Learning Materials**) are most learned in Africa region (**GeoLocation**)?
- etc.

There are several ways to render a diagram from a star schema data model. We use server-side web programming technology (Java Servlet) to retrieve data from the data model. We used HTML5, JavaScript, and CSS to build the interface, and implemented some third-party data visualization JavaScript libraries (HightChart, D3JS, AmCharts, etc.) to render some complex and beautiful charts in an instant.

However, most JavaScript libraries require data in JSON format to render charts. Therefore, the data must go through a series of transformations (from text files to database records to JSON). This is one of the typical responsibilities that data engineers always have to deal with from unstructured (text files) to structured (tabular records) and structured to semi-structured (JSON).

The purpose of the visualization is to highlight the user behavior patterns we discovered for further analysis by language experts. Typically, we use SQL statements to query the star schema data model and design charts that are best suited to emphasize the data schema when we find something interesting.

E. PREDICTIVE AI MODELS

In order to recommend useful Portuguese verbs for users to learn, we utilize different machine-learning approaches to build predictive models. We then evaluate each model to see which one works best for our situation.

1) PREDICTIVE MODELS BASED ON LSTM AND GRU

Since LLM requires an advanced and powerful GPU computing environment, which is not feasible for us, we believe that LSTM and GRU machine learning methods should be able to meet the needs. To prepare training data for LSTM and GRU networks, we convert the data in the data pond into text files. Each line of the file contains a data record in the format of an input verb followed by a suggested verb, separated by commas. The format of each line is as follows: $l_n = (v, s_1, s_2, \ldots)$

Once the data are ready, we start training the prediction model using LSTM and GRU unsupervised neural networks. However, after training for 100 epochs, the accuracy of LSTM can only reach 0.3417 and the loss is 2.8451, and the accuracy of GRU can only reach 0.3187 and the loss is 2.9272. The results were not satisfactory and we were disappointed.

2) PREDICTIVE MODELS BASED ON LINEAR REGRESSION

Students may have many different reasons for learning Portuguese verbs in groups. One obvious reason is the nature of the specific conjugation rules that verbs follow. In other words, Portuguese verbs with the same conjugation rules are always compared. Therefore, we designed to train the AI network using a linear regression model by adding an additional feature (conjugation rule). The new data format is amended as follows: $l_n = (v, r, s_1, s_2, \dots)$

Algorithm 1 is the pseudo-code for constructing a linear regression AI model. It is implemented using Python and the Scikit-learn library and is based on Géron's textbook [53].

After training the supervision network, we define two test datasets to evaluate the performance of the predictive model. The first test case (verb to suggestion verb) contains 100 verbs to evaluate the suggestion verb. The test results showed that the prediction model successfully predicted 87 related verbs and 13 irrelevant verbs. The second test case (verb-to-verb suggestion) contains 100 suggested verbs to evaluate the original verbs. The test results show that the prediction model successfully predicts 81 original verbs from the suggested verbs. All test results are summarized as a confusion matrix in Table 1.

Algorithm 1 Linear Regression Algorithm

```
Input: $f$                                        ▷ Input text file.
procedure CreateModel()
    $i \leftarrow 0$                              ▷ A counter.
    while $f(l) \neq EOF$ do                      ▷ Read each line of File.
        $v[i] \leftarrow l[0]$                    ▷ First column as verb.
        $r[i] \leftarrow l[1]$                    ▷ Second column as rule.
        $s[i] \leftarrow l[2]$                    ▷ Others as suggestion.
        $i \leftarrow i + 1$                      ▷ Increment counter.
    $X \leftarrow [v, r]$                         ▷ X as Input.
    $y \leftarrow s$                              ▷ y as Output.
    $model \leftarrow LinearRegression()$
    $model.fit(X, y)$
Input: $v, r$                                     ▷ Verb and rule number.
function PredictVerb()
    $s \leftarrow model.predict(v, r)$
    return $s$                                    ▷ Return suggesting verbs.
```

Although the results cannot achieve an accuracy of more than 90%, this prediction model can already anticipate useful verbs for students to learn relevant materials. Therefore, in this study, we choose the linear regression method to recommend Portuguese verbs to users.

IV. PRELIMINARY FINDINGS

The mobile app is published on the Apple Store and Google Play and can be installed in most countries. Since Google Play may not be accessible in some regions, the official website provides *Android Package Kit* (APK) files for Android users to download and install. After converting all possible log archives (including Apple Store and Google Play) between February 13, 2018, and June 30, 2023, we summarized some important statistics as shown in Table 2.

TABLE 1. Confusion matrix.

		Actual	
		Positive	**Negative**
Predict	**Positive**	87	13
	Negative	19	81

TABLE 2. Fundamental statistics on the usage of the mobile app.

Items	Values
Number of Dates	1,964*
Number of Hits	12,424,070
Number of Unique IP Address	185,048
Number of Visited Countries	148
Number of Visited Cities	694

* *2018-02-13 to 2023-06-30*

The star data model is the core of this study, which provides multi-dimensional information for our research. All visualization diagrams are based on information in this data model. In this section, we use the onion-peeling approach to evaluate and analyze all relevant statistics in this star schema data model. We calculate the number of hits across different dimensions (time, learning material, region, etc.) to drill down and discover any possible patterns.

A. USER DISTRIBUTION

Table 3 shows the most visited regions in the world. Some regions are not shown in the table because we wanted to focus on Portuguese-speaking regions.

Although Macao is part of China, its culture and education system are slightly different from mainland China. Therefore, even if Macao's population is less than 700,000, an independent analysis is necessary. In addition, since the mobile application was developed and launched in Macao, Macao is expected to become the region with the highest usage rate. As mentioned before, this mobile application is designed for people to learn Portuguese, so the regions in the ranking list (shown in Table 3) are mainly Portuguese-speaking regions.

Furthermore, according to the continental division, we can divide them into four groups (Africa, America, Asia, and Europe) to compare with Macao (the fifth group). Therefore, we only focus on the higher-ranking Portuguese-speaking regions in these five groups and do not study other insignificant data. The regions under study are listed below in

TABLE 3. Worldwide distribution of users of the app.

Rank	Regions*	Visits
1.	Macao SAR, China	5,785,193
2.	China	4,190,118
3.	Portugal	1,443,582
	...	...
5.	Brazil	169,633
6.	Angola	166,838
	...	...
8.	Mozambique	157,001
	...	...
17.	Cabo Verde	126,568
	...	...
21.	São Tomé and Príncipe	94,508
	...	...
46.	Guiné-Bissau	8,921
	...	...
50.	Timor-Leste	4,932
	...	...
66.	Guiné Equatorial	3,753

* *China and Portuguese-speaking countries/regions*

alphabetical order and all future studies will be based solely on these regions.

1) AFRICA

Angola (AO), Cape Verde (CV), Guiné Equatorial (GQ), Guiné-Bissau (GW), Mozambique (MZ), and São Tomé and Príncipe (ST).

2) AMERICA

Brazil (BR).

3) ASIA

China (CN), and Timor-Leste (TL).

4) EUROPE

Portugal (PT).

5) MACAO

Macao (MO).

B. TARGETED AUDIENCE

By analyzing the data in the star schema data model, we can render a visual diagram (shown in Fig. 2) to show the distribution of mobile app users around the world. Fig. 2 is a visual diagram called a heatmap, where red represents the highest frequency of access, gray represents the lowest frequency, and yellow represents the middle range. This map-like image uses areas as the smallest units to form a largely connected area, which can provide context for the user to perceive the map metaphor [54].

There are many reasons why a region ranks higher, and we have concluded some of the possibilities below. The region has Portuguese as its official language (Portugal, Brazil, Angola, Mozambique, etc.), has Chinese-Portuguese language schools (China, Macao, Portugal, etc.), and is a tourist region (Hong Kong, United States, Kingdom, Taiwan, etc.), and provides many proxy servers or VPNs for use (United States, France, etc.) [55].

Therefore, we rendered heatmap diagrams of the top countries to delve deeper, which are Angola in Fig. 3, Brazil in Fig. 4, China in Fig. 5, and Portugal in Fig. 6.

By the same token, all of these visual diagrams are heatmaps and show the most visited regions in red. The visual diagrams include some of the most visited provinces in each country. For example, Luanda in Angola, São Paulo in Brazil, Guangdong, Beijing, and Zhejiang in China, Lisbon, Porto, and Leiria in Portugal, etc. They may not be the capital of these countries, but there are many reputable higher education institutions offering courses related to Chinese and Portuguese. Some of these reputable schools are listed below.

1) LUANDA IN ANGOLA

- Universidade Agostinho Neto – Instituto Confúcio

2) SÃO PAULO IN BRAZIL

- Universidade de São Paulo – Bacharelado em Letras – Chinês
- Universidade Estadual Paulista – Instituto Confúcio

3) GUANGDONG, BEIJING, ZHEJIANG IN CHINA

- Sun Yat-sen University – School of Foreign Languages
- Guangdong University of Foreign Studies – Faculty of European Languages and Cultures – Portuguese
- Guangzhou Xinhua University – Portuguese
- Peking University – The Department of Spanish-Portuguese-Italian Languages and Literatures
- Zhejiang International Studies University – School of European Languages and Cultures-Portuguese

4) LISBON, PORTO, LEIRIA IN PORTUGAL

- Universidade de Lisboa – Instituto Confúcio
- Universidade do Porto – FLUP – Chinês
- Instituto Politécnico de Leiria – Licenciatura em tradução e interpretação Português-Chinês / Chinês-Português

Consequently, regions with more language schools have higher hit rates and the visual diagrams will turn red. After the mobile app was released in 2018, our teachers and students promoted and spread the word to all language schools around the world. As shown in the image above, we strongly believe that language school students (our target group) are the majority of users. Our users fit neatly into these five groups, and we are excited to see this mobile app serve in this way.

C. DEVICES AND PLATFORMS

By using the device dimension of the star schema data model, we can consolidate the mobile platform usage of users in different regions, as shown in Fig. 7.

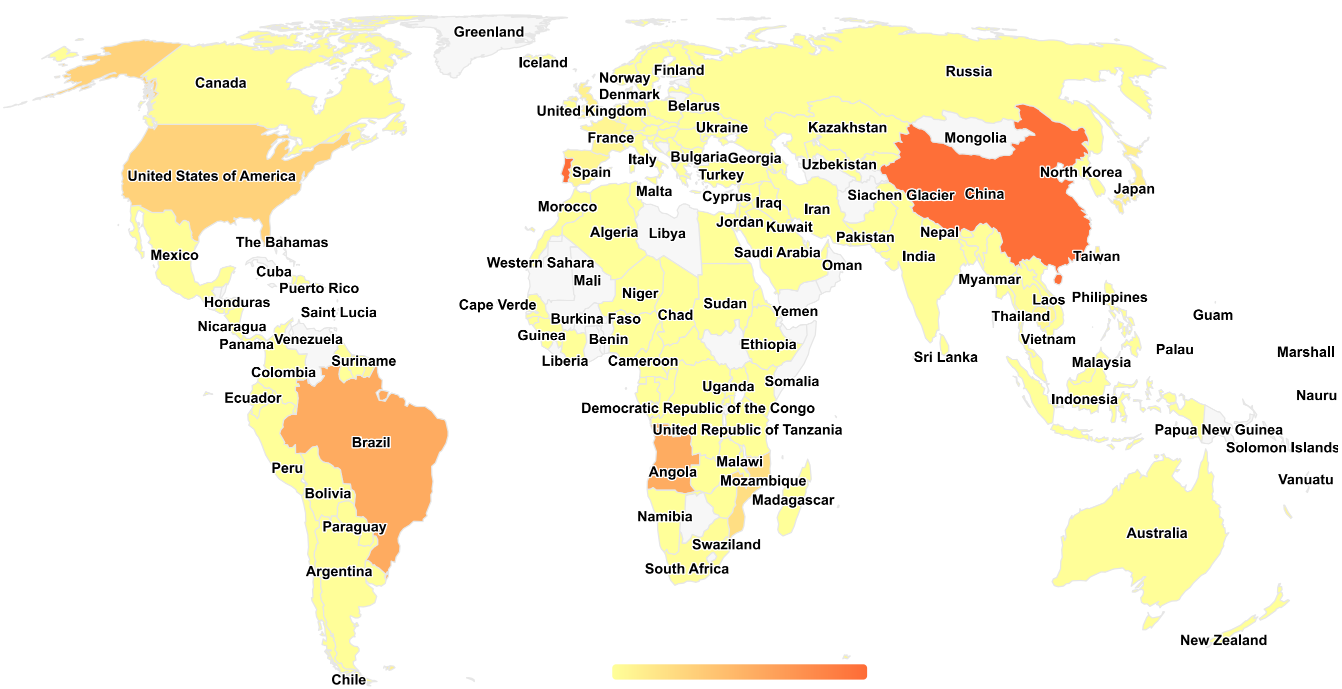


FIGURE 2. Heatmap diagram of the worldwide distribution of users of the mobile app.

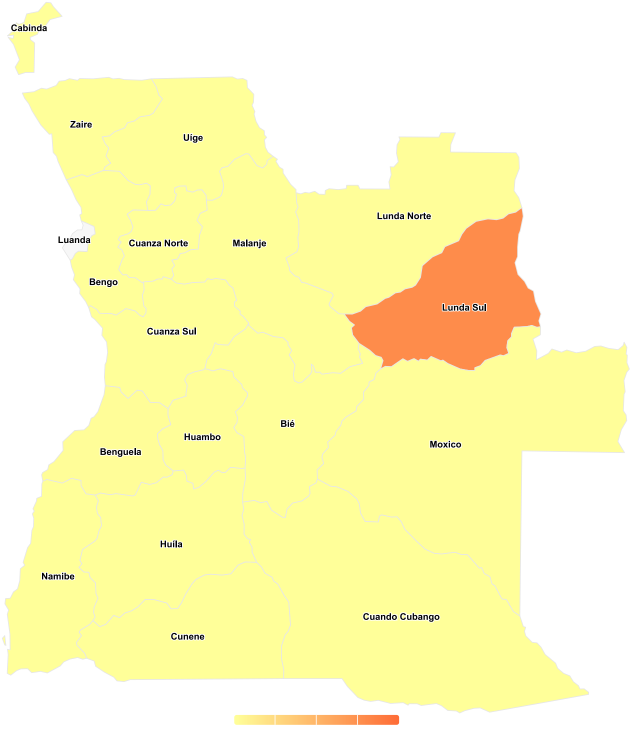


FIGURE 3. Heatmap diagram of user distribution in Angola of the app.

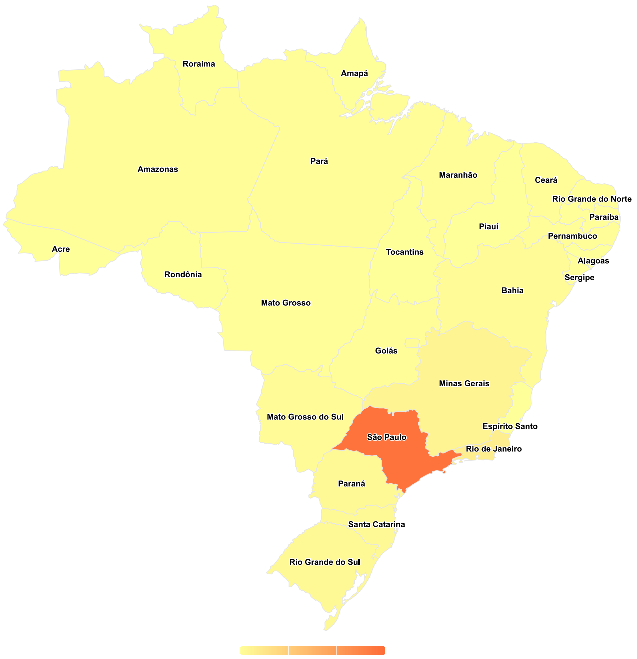


FIGURE 4. Heatmap diagram of user distribution in Brazil of the app.

Among them, 72.47% of people use the iOS platform, and only 27.53% use the Android platform. In general, our target audience (learning Portuguese students) is more likely to use devices running on the iOS platform. Moreover, only Mozambique and Timor-Leste have more Android users than iOS. Angola and Brazil have similar numbers of users on both platforms. Other regions are dominated by the iOS platform. Looking into more details about iOS users in our data model,

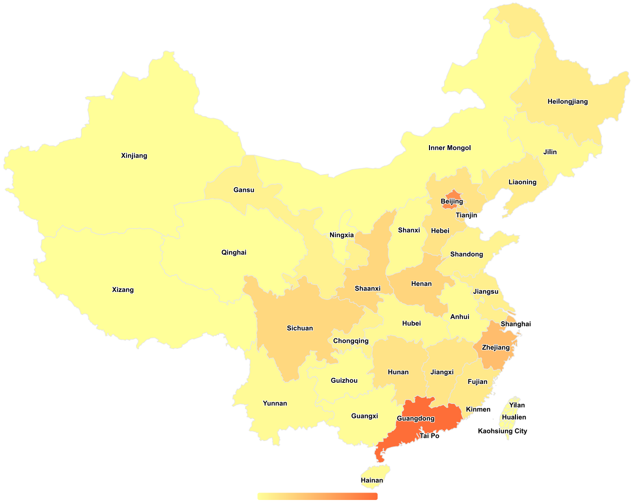


FIGURE 5. Heatmap diagram of user distribution in China of the app.

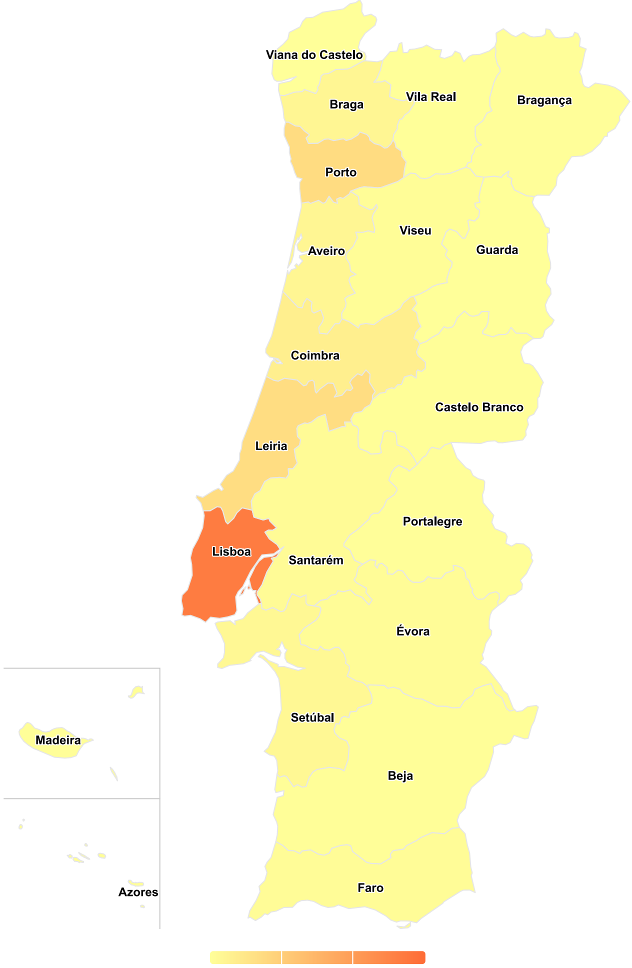


FIGURE 6. Heatmap diagram of user distribution in Portugal of the app.

approximately 88% are iPhone users and only 12% are iPad users. Consequently, we have more iPhone users than others.

According to recent "Global Smartphone Market Share" statistics [56], [57], Android users (approximately 79%) dominate iOS users (approximately 21%) in the market. Although China's smartphone market share is led by Samsung (22.5%) and Apple (20.5%), Apple still has a huge user base in China. According to statistics, Apple sells more iPhones in China than in the United States [56]. Therefore, China is one of Apple's largest markets and we expect there will be more iOS users in China.

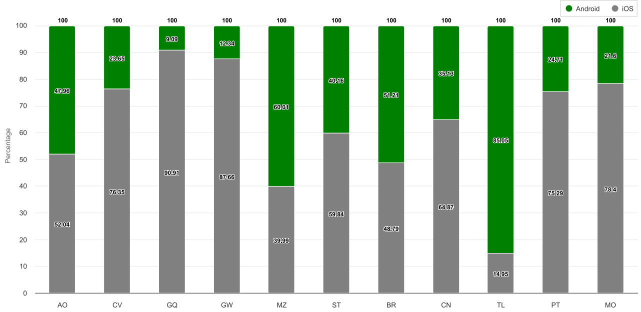


FIGURE 7. Market share of mobile platforms in different regions.

Access to certain Google services is restricted in China, and people may not be able to install all the apps available on the Google Play Store. Huawei, one of China's largest telecommunications companies, has been banned by the U.S. government, severely affecting its global market share [58]. Therefore, people are willing to use iOS devices to explore new applications and learn new things in advance.

The expensive iPhone has a special status and symbol for the Chinese. Owning an iPhone represents a wealthy life and can improve your social status, especially among young people. According to our statistics, the hardware models and platform versions of the devices are relatively new, and people are constantly updating their devices. As discussed earlier, most of our users are language students using iPhones, which may explain why we have a lot of iOS users, as students are a younger user base willing to explore new horizons.

Our data model revealed the fact that iPhones are a trend among students learning Portuguese (our target group). Therefore, we can recommend that people use the iPhone as a development platform when creating and updating new educational software, and pay attention to new features of iOS to improve user experience. As a result, we can apply the concepts of MALL and SDT to develop educational mobile apps.

D. USAGE OVER TIME

By including the time dimension in the data model, we can analyze the access time behaviors of our users. Table 4 summarizes the monthly number of hits from February 2018 to June 2023.

These numbers include all clicks when users access the app. Fig. 8 visualizes the information in Table 4. This line chart makes it easy to form four patterns across time periods (x-axis).

TABLE 4. Number of hits on the app over different time periods.

Month	Hits	Month	Hits	Month	Hits
2018-01	-	2019-01	294,401	2020-01	73,763
2018-02	472,964	2019-02	236,491	2020-02	101,120
2018-03	410,878	2019-03	480,467	2020-03	158,106
2018-04	337,822	2019-04	387,624	2020-04	211,998
2018-05	296,145	2019-05	401,825	2020-05	165,602
2018-06	220,167	2019-06	286,413	2020-06	140,085
2018-07	167,369	2019-07	296,791	2020-07	86,555
2018-08	100,993	2019-08	282,937	2020-08	50,344
2018-09	273,097	2019-09	**577,387**	2020-09	103,207
2018-10	382,520	2019-10	**685,180**	2020-10	125,310
2018-11	389,243	2019-11	172,404	2020-11	119,617
2018-12	287,882	2019-12	106,277	2020-12	91,092
2021-01	85,851	2022-01	66,173	2023-01	205,525
2021-02	58,674	2022-02	55,290	2023-02	252,564
2021-03	123,942	2022-03	106,567	2023-03	355,565
2021-04	118,632	2022-04	89,993	2023-04	254,492
2021-05	107,725	2022-05	88,525	2023-05	204,179
2021-06	81,147	2022-06	74,596	2023-06	67,334
2021-07	51,805	2022-07	49,100		
2021-08	52,937	2022-08	45,697		
2021-09	102,064	2022-09	89,698		
2021-10	104,708	2022-10	93,016		
2021-11	117,194	2022-11	98,197		
2021-12	81,027	2022-12	167,777		

* *statistics from February 2018 to June 2023*

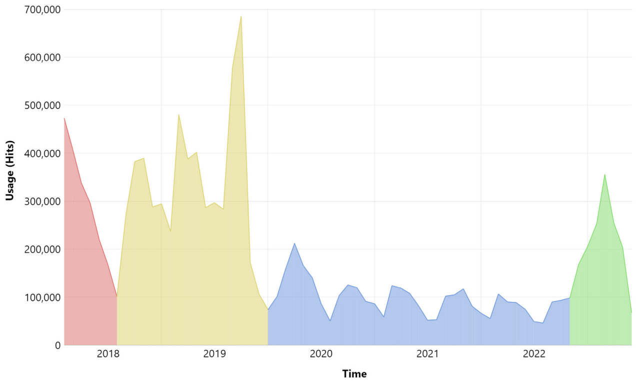


FIGURE 8. Number of hits on the app over different time periods.

The red pattern shows the decline from February to August 2018. The mobile app was released on February 13, 2018, and people are excited to use it to explore all the new features at that time. However, as time went by, the enthusiasm slowly faded, and the hit rate of the app continued to decline.

As the new semester began in September 2018, hit rates began to increase as teachers used the app more frequently in their courses. The yellow pattern shows that the mobile app became popular due to its highest heap records set in September and October 2019. The waveform shows that peaks occur during semesters (March, April, May, September, October, etc.) and low peaks occur during holidays (July, August, December, etc.). This pattern of fluctuations happens to follow the school year in a reasonable manner.

TABLE 5. Number of hits for different learning materials.

Learning Materials	Hits	Percentage
Verb	6,051,643	48.71%
Daily Word	2,336,820	18.81%
Conversation	2,167,315	17.44%
Vocabulary	1,203,605	9.69%
Video	196,397	1.58%
Quiz	131,866	1.06%
Others	336,424	2.71%
Total:	12,424,070	100.00%

The blue pattern reveals the difficult times that most people have been going through recently. A well-known disease (COVID-19) started in December 2019 and lasted until 2023, affecting us severely [1], [59]. All normal activities are closed and people are forced to stay at home. As a result, many behaviors change accordingly. People need to work from home and schools have moved to online classes. Overall hit rates for mobile apps also declined during this period. By the same token, a wavy pattern is clearly formed, indicating that the hit rate is mainly affected by the school year.

Finally, a green pattern indicates that hit rates have returned to normal as the COVID-19 pandemic declares an end to the public health emergency of international concern [1], [59]. Although we only have data for the first half of 2023, the hit rate can reach 2019 levels.

Taken together, the overall results are in line with our expectations. It shows how mobile app usage varies by academic year. We were able to cluster four patterns across different time periods, with the COVID-19 pandemic being one of the major factors affecting hit rates from 2020 to 2022.

V. DATA ANALYSIS

The previous section "Preliminary findings" provided insight into how our mobile applications are used. This section yields a more in-depth study aimed at contributing more useful information when developing educational applications.

A. LEARNING MATERIALS

Fig.8 is based on all hit rates combined together for each click by a user. To understand which learning materials are more popular, we added the "Learning Materials" dimension to drill into statistical values for analysis. Table 5 summarizes the number of hit counts on the mobile application learning materials.

1) MOST FREQUENTLY USED LEARNING MATERIALS

Apparently, people mainly use this mobile app to learn Portuguese verb conjugations. Portuguese verb conjugations are very complex. There are over 100 different sets of rules for different verbs, and each verb has around 70 conjugations [43]. Memorizing all of this correctly is not an easy task.

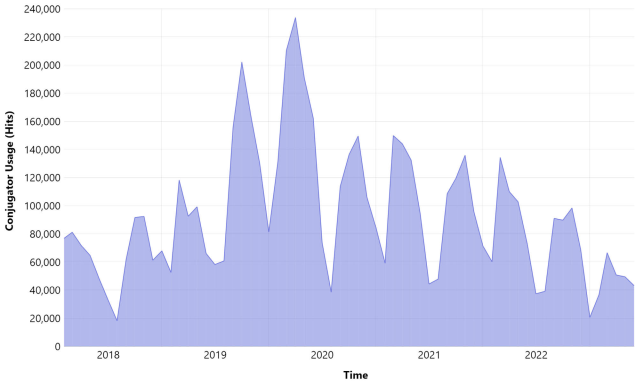


FIGURE 9. Number of hits of verbs in different time periods.

TABLE 6. Usage distribution of learning materials in different regions.

Regions*	Verb	Conv.	Daily	Vocab.	Others
MO	41.13%	23.11%	19.33%	12.18%	4.25%
CN	62.23%	13.40%	13.08%	8.67%	2.62%
PT	46.78%	17.06%	24.01%	9.71%	2.44%
BR	36.57%	22.17%	25.93%	11.28%	4.05%
AO	28.35%	26.23%	25.24%	15.20%	4.98%
MZ	33.40%	26.71%	21.49%	13.68%	4.72%
CV	30.85%	30.04%	18.12%	16.12%	4.87%
ST	28.45%	26.50%	28.09%	12.42%	4.54%
GW	10.05%	32.47%	30.41%	21.91%	5.16%
TL	17.59%	31.48%	28.70%	14.35%	7.88%
GQ	45.30%	14.80%	35.40%	2.70%	1.80%

* *ordered according to Table 3*

Therefore, being able to display all conjugated forms of a verb is very useful for people learning Portuguese.

Fig. 9 provides a visual representation of verb learning over time. If we compare Fig. 8 with Fig. 9, even during the COVID-19 pandemic, people are still using mobile apps to learn conjugations of verbs. In fact, people are spending more time learning to conjugate verbs than before the COVID-19 pandemic, especially as schools began to close in 2020. Furthermore, the highest peak occurred in April 2020.

Therefore, the design of this mobile app can fully implement the main ideas of MALL and SDT, allowing students to learn independently without relying on traditional classrooms, and successfully transform from teacher-centered to learner-centered.

2) LEARNING MATERIALS USAGE IN DIFFERENT REGIONS

We know that query verbs are more common than other learning materials, but is this the same across the five different groups? Table 6 summarizes the use of learning materials in different regions.

For most people who use this mobile app to learn verb conjugations, the results are expected. However, we cannot ignore the fact that some people in Africa (Guiné-Bissau, Cape Verde, Mozambique, etc.) like to learn conversation. Therefore, mobile app developers need to create more conversations to maintain hit rates.

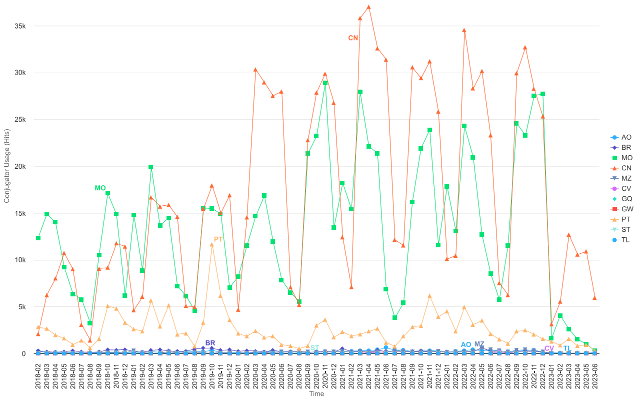


FIGURE 10. Number of hits of verbs in different regions over different time periods.

B. HOW VERBS ARE BEING LEARNED?

Fig. 10 visualizes the number of hits over time for verbs in different regions. The two overwhelming lines in the diagram are China in red and Macao in green. In fact, China's red line is even higher than Macao's green line, because according to Table 3 and Table 6, the proportion of China (> 2.6 million hits) is 62.23%, which is slightly higher than Macao ($\simeq$ 2.38 million hits) at 41.13%.

Results showed that people in mainland China and Macao use this mobile app to learn Portuguese verbs regularly. The other lines at the bottom and Portugal's orange line are statistically trivial. The fluctuations in Fig. 10 are similar to Fig. 9, with high peaks occurring during semesters and low peaks occurring during holidays. Additionally, the red line shows people using more during the COVID-19 pandemic, which is the likelihood of classes moving online. The sudden drop in numbers in December 2022 and January 2023 came as many cities in China returned to normal and claimed the COVID-19 pandemic was no longer the biggest threat to people.

Therefore, if we wanted to further study how people learn verbs in this mobile app, we could just focus on China and Macao. We believe these two regions represent the majority.

C. MOST LEARNED VERBS

The mobile app contains over 15,000 verbs and over 100 different verb conjugation rules. Each rule contains different groups (Modo Indicativo, Condicional, Modo Conjuntivo, Modo Imperativo), different tenses (Presente, Pretérito imperfeito, Pretérito perfeito, Pretérito Mais-que-perfeito, Futuro, etc.), and different pronouns (eu, tu, você, nós, vós, etc.), with over 70 different forms of rules. Even for natives, mastering the conjugations of Portuguese verbs fluently can be a tough job. Since Chinese verbs do not have various forms, we believe that the conjugation of Portuguese verbs should be one of the most difficult problems for Chinese people to overcome when learning Portuguese.

Table 7 provides an idea of how Portuguese verbs are conjugated in different circumstances [60]. Some people in

TABLE 7. Conjugation forms of "VER".

Modo Indicativo					
Presente		Pretérito imperfeito		Pretérito perfeito	
eu	vejo	eu	via	eu	vi
tu	vês	tu	vias	tu	viste
ele/ela/você	vê	ele/ela/você	via	ele/ela/você	viu
nós	vemos	nós	víamos	nós	vimos
vós	vedes	vós	víeis	vós	vistes
eles/elas/vocês	vêem	eles/elas/vocês	viam	eles/elas/vocês	viram
Modo Indicativo				**Condicional**	
Pretérito mais-que-perfeito		Futuro		Presente	
eu	vira	eu	verei	eu	veria
tu	viras	tu	verás	tu	verias
ele/ela/você	vira	ele/ela/você	verá	ele/ela/você	veria
nós	víramos	nós	veremos	nós	veríamos
vós	víreis	vós	vereis	vós	veríeis
eles/elas/vocês	viram	eles/elas/vocês	verão	eles/elas/vocês	veriam
Modo Conjuntivo					
Presente		Pretérito imperfeito		Futuro	
eu	veja	eu	visse	eu	vir
tu	vejas	tu	visses	tu	vires
ele/ela/você	veja	ele/ela/você	visse	ele/ela/você	vir
nós	vejamos	nós	víssemos	nós	virmos
vós	vejais	vós	vísseis	vós	virdes
eles/elas/vocês	vejam	eles/elas/vocês	vissem	eles/elas/vocês	virem
Modo Imperativo				**Infinitivo pessoal**	
Afirmativo		Negativo			
—		—		eu	ver
tu	vê	tu	não vejas	tu	veres
ele/ela/você	veja	ele/ela/você	não veja	ele/ela/você	ver
nós	vejamos	nós	não vejamos	nós	vermos
vós	vede	vós	não vejais	vós	verdes
eles/elas/vocês	vejam	eles/elas/vocês	não vejam	eles/elas/vocês	verem
Infinitivo impessoal		**Gerúndio**		**Particípio passado**	
ver		vendo		visto	

* *In the future conjunctive, the verb "ver" can have a conjugation that is identical to the infinitive of the verb "vir"*

science and engineering fields may think that the periodic table of chemical elements is easier to memorize. One might say that among these 15,000+ verbs, there are many rare verbs, but there are still hundreds of common verbs used in our daily lives. Nevertheless, the verb "ver" in Table 7 is one of the irregular verbs, and its conjugations are more complicated than those of regular verbs.

If we want to know exactly which verbs people are learning, we can query based on the time, material, and geographical location dimensions of the star data model. As mentioned previously, we can only focus on China and Macao because of their size. Fig. 11 and Fig. 12 below are intended to show the most commonly learned verb conjugations.

These hierarchical diagrams are like a blossom with many branches and sub-branches. They mainly consist of three levels of nodes (small circles). The root node in the middle of the diagram represents a specific region (China or Macao), each branch node (connected to the root node) represents the rule of the verb, and the sub-branch nodes or end nodes represent the verb.

In addition, the size of the node represents the frequency of learning, and we use red to indicate high frequency. There is a large group of nodes on the left side of Fig. 11, which is the rule followed by most Portuguese verbs. Approximately 43.3% (6,500 / 15,000) of verbs follow the "-ar" radical rule ("falar", "passar", "trabalhar", etc.). In fact, only rules with regular verbs can have clusters of this size. Therefore, mainland Chinese learn a wider range of regular verbs.

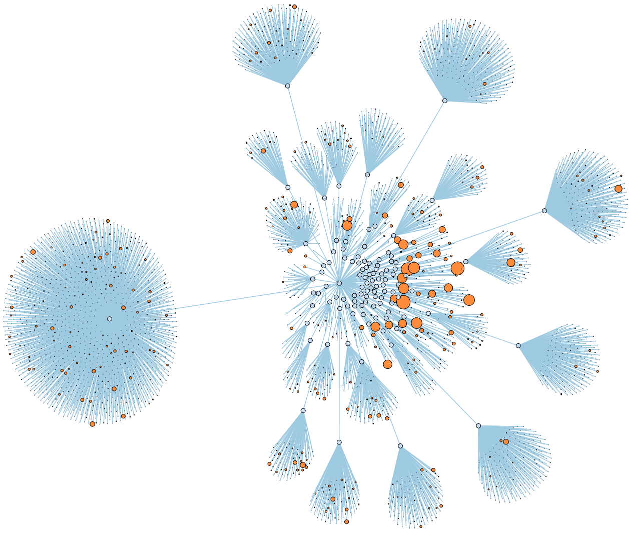


FIGURE 11. Most learned verbs in China.

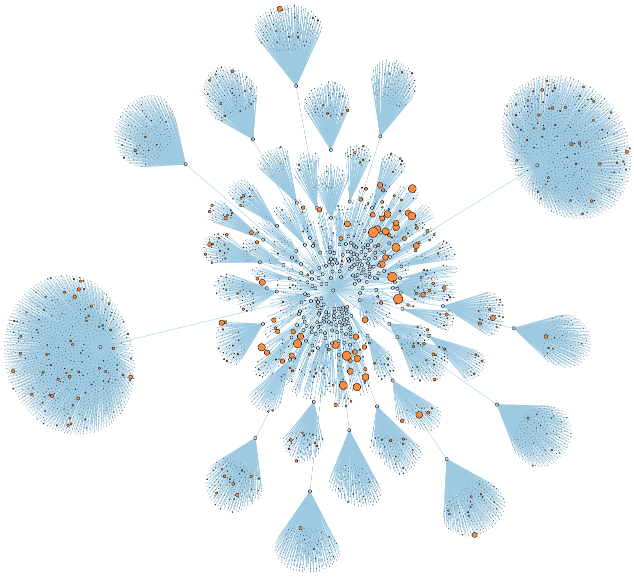


FIGURE 12. Most learned verbs in Macao.

The larger size red nodes are usually distributed in the middle of the two diagrams, and they are the most learned verbs. A closer look shows that they only have one branch and one sub-branch, and we call them "singletons". This singleton pattern means there is only one instance of a type [61]. In other words, a rule is followed by only one verb.

In reality, these singletons are rarely used verbs with their own unique rules. Sometimes, even native Portuguese speakers cannot make them crystal clear because they are mostly irregular verbs. There are about 25 singletons and about 50 rules that are followed by less than 5 verbs in the mobile app. Listed below are some verbs that can render big red nodes.

autodefender-se, coerir, concernir, dar, delinquir, desmilinguir-se, embair, estar, fazer, frigir, haver, ir, moscar, perder, pôr, poder, precluir, reaver, requerer, ressarcir, retorquir, reunir, sair, ser, servir, sortir, ter, trazer, ver, vir, etc.

Because of their rarity, our users really pay more attention and learn them over and over again. This behavior causes some large red nodes to appear in the middle of the graph, forming a singleton pattern.

If we compare the two diagrams, Macao has two large node clusters, which are regular verbs under the two "-ar" radical rules ("falar", "passar", "trabalhar", etc.) and ("barbear", "chatear", "passear", etc.). To distinguish between these two large clusters, verbs with the ordinary "-ar" suffix are on the left, and verbs with the "-(e)ar" suffix are on the right. The latter verb display a distinctive trait: the vowel 'e' is replaced by the diphthong 'ei' in the stressed syllable [60], [62]. Although they are both regular "-ar" radical verbs, the difference in vowels can be confusing to our Portuguese language learners. Therefore, one must be able to classify verbs to follow the "-(e)ar" or normal "-ar" rules.

Moreover, there are no big red nodes in these two clusters like those singletons, indicating that they are not as difficult as those irregular verbs in the middle of the diagram. We believe that the similarity in size of these two large clusters indicates that people always compare them to different verbs. Therefore, unlike those who spend a longer time learning irregular verbs, one only needs to check whether the verb follows the "-ar" or "-(e)ar" suffix rules.

Furthermore, Fig. 12 has more small node clusters, which means that Macao people learn more different rules than mainland Chinese people. Since Portuguese is one of the official languages of Macao, Macao natives may need a more comprehensive study of all verbs. On the other hand, mainland Chinese will pay more attention to those rare verbs with larger red nodes. Therefore, this explains the fact that there are more hits but fewer node clusters in Fig. 11. Specifically, there are approximately 3,700+ verbs in Fig. 11, 4,000+ verbs in Fig. 12, and Macao has 300 verbs more.

The beauty of data visualization is the ability to pack many dimensions into a two-dimensional medium [63], and using modern technology to display as much information as possible in a single diagram is an interesting problem.

D. MOST LEARNED VERB GROUP IN DIFFERENT REGIONS

We know there is an increasing emphasis on learning irregular Portuguese verbs. However, what are these irregular verbs? Do people in different regions learn the same irregular verbs? To answer these questions, we need to ask the star schema data model again.

Table 8 summarizes the most learned verbs in China and Macao, and Table 9 summarizes the most learned verbs in Portugal, Brazil, and Angola.

TABLE 8. Most learned verbs in China and Macao.

Rank	China		Macao	
1.	vir	4.7606%	ir	4.2362%
2.	ver	4.4642%	vir	3.4534%
3.	ir	4.2497%	ver	3.3236%
4.	ter	3.2147%	ter	3.0395%
5.	fazer	3.0274%	fazer	2.9212%
6.	dar	3.0003%	ser	2.8078%
7.	ser	2.7613%	estar	2.4479%
8.	sair	2.3867%	dar	2.2043%
9.	pôr	2.3401%	poder	1.9935%
10.	poder	2.3109%	sair	1.9533%

* *regions ordered by total number of hits*

TABLE 9. Most learned verbs in portugal, brazil, and angola.

Rank	Portugal		Brazil		Angola	
1.	vir	4.73%	ir	4.00%	vir	4.28%
2.	ver	4.04%	vir	3.93%	ver	3.74%
3.	ir	3.82%	ver	2.87%	ir	3.43%
4.	ter	3.01%	fazer	2.80%	fazer	2.75%
5.	fazer	2.95%	ser	2.80%	ser	2.50%
6.	dar	2.93%	ter	2.59%	ter	2.19%
7.	ser	2.70%	estar	2.35%	dar	2.03%
8.	estar	2.21%	dar	2.14%	poder	1.94%
9.	pôr	2.17%	sair	1.46%	estar	1.89%
10.	poder	2.14%	poder	1.44%	sair	1.60%

* *regions ordered by total number of hits*

According to previous statistics, Portuguese verb learning has received approximately 2.6 million hits in China and approximately 2.38 million hits in Macao. Among them, more than 4% of the hits in China (i.e. 100,000+ hits) are learning the verbs "vir", "ver", and "ir". Interestingly, the top three verbs in Macao are exactly the same as those in China. Moreover, if we compare the top ten verbs of Table 8, they are almost identical except for a slightly different order.

By the same token, we used the star schema data model to calculate the most learned verbs in other regions, and the results are summarized in Table 9. The top 10 verbs are "coincidentally" the same in Portugal, Brazil, and Angola but in a slightly different order. This "coincidence" shows that these Portuguese verbs are difficult to figure out. Therefore, no matter where in the world you learn these verbs, you need to learn them over and over again.

The statistical values of the five representatives (CN, MO, PT, BR, and AO) are listed in Table 8 and Table 9. We also queried other regions (CV, GQ, GW, MZ, ST, and TL) of the five groups and obtained similar results. While the ten verbs may not be the same, the percentage distributions are pretty much alike. Moreover, these "new" verbs (dizer, querer, saber, etc.) from other regions are actually ranked between 11th and 20th in Table 8 and Table 9.

TABLE 10. Most learned verbs by their hit rates.

Levels	Verbs	Range
1	ir, ver, vir	$> 3\%$
2	fazer, ser, ter	$2.5 \sim 3\%$
3	dar, estar	$\simeq 2.5\%$
4	poder, pôr, sair	$2 \sim 2.5\%$
5	others	$< 2\%$

TABLE 11. Most learned verbs in groups.

Rank (n)	Verb Group $g(n)$	Frequency $f(g)$
1.	ver, vir	1,362
2.	ir, vir	535
3.	abraçar, abrir	453
4.	estar, ser	265
5.	ir, ser	244
6.	ser, ter	201
7.	ler, ver	198
8.	ir, ver	190
9.	ir, ver, vir	181
10.	ir, ter	150
11.	pôr, vir	143
12.	dar, vir	140
13.	dar, ir	138
14.	fazer, ir	127
15.	estar, ter	126
16.	abraçar, acabar	114
17.	sair, vir	112
18.	dar, ver	108
19.	ter, ver	108
20.	abar, abraçar	108
21.	ter, vir	104
22.	poder, ter	102
23.	fazer, ver	101

* *verb groups with more than 100 visits*

Consequently, we classify the top 10 verbs in Table 8 and Table 9 into different levels according to the average hit rate, as shown in Table 10. Obviously, Level 1 is the favorite of all Portuguese learners, and people will definitely pay a lot of attention to the verbs that go on to Level 1. Although Portuguese language learners are aware of and understand them well, they often need to double-check their conjugation forms.

E. VERBS IN GROUPS

On the one hand, the two large node clusters in Fig. 12 indicate that there are many similar verbs that may confuse us. On the other hand, the top ten verbs in Table 10 can be classified into four groups. Hence, do people usually learn verbs together to make comparisons? By using user sessions as mentioned in Section III-D, we can study the frequency of verbs in groups. Table 11 summarizes the most frequently learned verbs in the group with more than 100 user session visits.

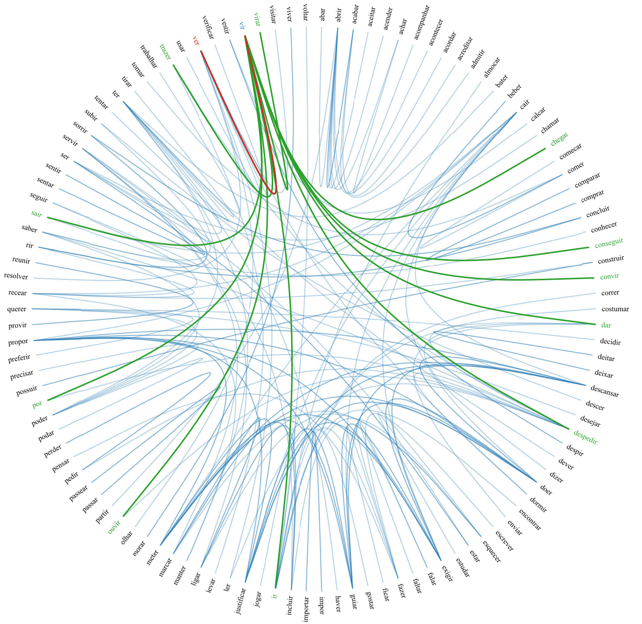


FIGURE 13. Correlation of most learned verbs.

The results show that $g(1) = \{ver, vir\}$ is the most learned verb group. The two verbs in the first group are very similar, but they have completely different meanings ("ver" means "to see") and ("vir" means "to come"). Furthermore, whenever we need to conjugate the first or third-person singular form of the verb "ver" into the future tense, we must use the word "vir", as shown in Table 7. Then, some people suggest continuing to use "ver" instead of "vir" to avoid conjugating a particular verb into other verbs. Yet these two verbs continue to confuse us over the years.

Verb groups can vary in size (approximately 2 to 10 verbs), as listed in Table 11, but most verb groups have only two verbs. It shows that people do not learn many verb conjugations in one session. We argue that people are more likely to learn both verbs simultaneously, primarily for comparison purposes.

The highest-ranked group, $g(1)$, has 1,362 visits, which is much higher than the other groups. If we look closely at other verb groups, it is easy to see that the "ver", "vir", and "ir" verbs are often combined with other verbs to form a group. It turns out that people always compare other verbs to these three verbs. Therefore, they are the most confusing verbs when learning Portuguese. Teachers should devise a novel mechanism to make students aware of their variations.

Some verb groups in the list are very different from others, they are the $g(3)$, $g(16)$, and $g(20)$ verb groups. Since all Portuguese verbs are listed alphabetically on the mobile app, the first two verbs happen to be "abraçar" and "abrir" on the "Verbos Frequentes" page. Therefore, the way learning materials are presented can also affect hit rates. Mobile app designers should always place the most important content prominently for the convenience of users.

By using the resulting data in Table 11, we can draw a circular diagram as shown in Fig. 13. It is designed to show the correlation of the most commonly learned verbs. The edges of the circle diagram are labeled with verbs, and in the case of verb groups, the edges are connected by blue lines. In addition, if you focus on an edge (verb), it turns its connecting lines and edges green to show related groups and red to show the most frequent groups.

In Fig. 13, the most coherent verbs are "ver" and "vir". It shows that they occur the most among all verb groups and can be expressed as Equation (4).

As mentioned before, about 43.3% of verbs follow the radical "-ar" rule, but people learn radical "-ir" verbs more frequently and learn them as a group in user sessions. This is because most "-ir" verbs are mostly irregular, and the way they are conjugated varies greatly. The results show that when people learn Portuguese verbs, they always learn related verbs as a whole. Therefore, we can provide an interface that puts together suggested verbs based on the verb groups we find for the convenience of the user.

VI. CONCLUSION

This study of the "Diz lá!" mobile app comprehensively examines its usefulness as a Portuguese language learning tool, particularly among Chinese speakers. Our multi-dimensional data model, enriched by five years of user behavior, is a wellspring of invaluable insights. This study contributes to our understanding of how MALL and SDT are implemented and received in the field, underscoring the pivotal role of MALL and SDT in promoting autonomous language learning, particularly during disruptions like the COVID-19 pandemic, thereby offering a distinctive perspective on evaluating the intersection of language learning, pedagogy, and technology.

Visual diagrams help us discover hidden information. Some interesting explorations are summarized below.

1) Users come predominantly from countries/regions with a high concentration of schools offering Portuguese language programmes.
2) Portuguese verb conjugations are the most popular learning material, and despite an overall decrease in app use, there has been increased interest in these materials during the COVID-19 pandemic.
3) Mainland Chinese speakers showed a more targeted approach to verb learning than Macau speakers, with a particular focus on irregular verbs.
4) Regular verbs with the "-ar" and "-(e)ar" suffixes are frequently compared, especially by learners in Macao.
5) The most commonly learned verbs were "ir", "ver", and "vir", indicating these verbs may pose specific challenges or areas of interest.
6) Learning verbs in groups is a common practice, and the most learned verb group is $\{ver, vir\}$.

Based on the results we found, we demonstrate an approach that uses modern machine learning methods to intelligently recommend relevant learning materials to users. As a result, the mobile app is able to anticipate some Portuguese verbs for users to conduct comparative studies. Furthermore, by using similar machine learning methods, all the findings in this article can also become new features for educational mobile apps.

These results highlight the adaptability of technology-facilitated language learning. Observing learner trends, such as the apparent focus on verb conjugations, can inform future MALL mobile app development and refinement. Moreover, this study sheds light on the differences in learning behaviors based on geographic location, which can help create a personalized learning experience.

This research establishes a dialogue between technology and language learning, emphasizing how the two can work together to optimize learning outcomes. The insights garnered can guide educators and software developers in creating personalized, effective learning solutions.

Our study's extensive and multi-faceted dataset provides a solid foundation for future research. Looking forward, this research study offers multiple avenues for future exploration. Exploring supplementary dimensions, such as user feedback, and harnessing artificial intelligence for predictive modeling may yield deeper insights into language learning behavior.

REFERENCES

[1] WHO. (2023). *Coronavirus Disease 2019 (COVID-19)*. [Online]. Available: https://www.who.int/health-topics/coronavirus/

[2] A. Kukulska-Hulme, "Mobile-assisted language learning," in *He Encyclopedia of Applied Linguistics*. Hoboken, NJ, USA: Wiley, Jun. 2020, pp. 1–9.

[3] G. Stockwell, *Mobile Assisted Language Learning: Concepts, Contexts and Challenges (Cambridge Applied Linguistics)*. Cambridge, U.K.: Cambridge Univ. Press, Jan. 2022.

[4] J. Burston, "MALL: The pedagogical challenges," *Comput. Assist. Lang. Learn.*, vol. 27, no. 4, pp. 344–357, May 2014.

[5] A. Kukulska-Hulme, "How should the higher education workforce adapt to advancements in technology for teaching and learning?" *Internet Higher Educ.*, vol. 15, no. 4, pp. 247–254, Oct. 2012.

[6] O. Viberg and Å. Grönlund, "Cross-cultural analysis of users' attitudes toward the use of mobile devices in second and foreign language learning in higher education: A case from Sweden and China," *Comput. Educ.*, vol. 69, pp. 169–180, Nov. 2013.

[7] H. C. Chen and R. F. Chung, "Interlanguage analysis of phonetic timing patterns by Taiwanese learners," *Concentric, Stud. Linguistics*, vol. 34, no. 1, pp. 79–108, Feb. 2008.

[8] H. Kim and Y. Kwon, "Exploring smartphone applications for effective mobile-assisted language learning," *Multimedia-Assist. Lang. Learn.*, vol. 15, no. 1, pp. 31–57, Apr. 2012.

[9] A. Kukulska-Hulme, "Language learning defined by time and place: A framework for next generation designs," in *Left to My Own Devices: Learner Autonomy and Mobile Assisted Language Learning*. New York, NY, USA: Emerald, Jan. 2012, pp. 1–13.

[10] W.-H. Wu, Y.-C. Jim Wu, C.-Y. Chen, H.-Y. Kao, C.-H. Lin, and S.-H. Huang, "Review of trends from mobile learning studies: A meta-analysis," *Comput. Educ.*, vol. 59, no. 2, pp. 817–827, Sep. 2012.

[11] C.-K. Chang and C.-K. Hsu, "A mobile-assisted synchronously collaborative translation–annotation system for English as a foreign language (EFL) reading comprehension," *Comput. Assist. Lang. Learn.*, vol. 24, no. 2, pp. 155–180, Apr. 2011.

[12] T. de Jong, M. Specht, and R. Koper, "A study of contextualised mobile information delivery for language learning," *Innov. Designing Mobile Learn. Appl.*, vol. 13, no. 3, pp. 110–125, Jul. 2010.

[13] N. A. Gromik, "Cell phone video recording feature as a language learning tool: A case study," *Comput. Educ.*, vol. 58, no. 1, pp. 223–230, Jan. 2012.

[14] Y.-M. Huang, Y.-M. Huang, S.-H. Huang, and Y.-T. Lin, "A ubiquitous English vocabulary learning system: Evidence of active/passive attitudes vs. usefulness/ease-of-use," *Comput. Educ.*, vol. 58, no. 1, pp. 273–282, Jan. 2012.

[15] D. V. Kolesova, L. V. Moskovkin, and T. I. Popova, "Urgent transition to group online foreign language instruction: Problems and solutions," *Electron. J. e-Learn.*, vol. 19, no. 1, pp. 21–41, Apr. 2021.

[16] J. Guo, F. Huang, Y. Lou, and S. Chen, "Students' perceptions of using mobile technologies in informal English learning during the COVID-19 epidemic: A study in Chinese rural secondary schools," *J. Pedagogical Res.*, vol. 4, no. 4, pp. 475–483, Oct. 2020.

[17] Z. Chen, W. Chen, J. Jia, and H. An, "The effects of using mobile devices on language learning: A meta-analysis," *Educ. Technol. Res. Develop.*, vol. 68, no. 4, pp. 1769–1789, Jun. 2020.

[18] F. Li, "Student and language teacher perceptions of using a WeChat-based MALL program during the COVID-19 pandemic at a Chinese University," *Educ. Sci.*, vol. 13, no. 3, p. 236, Feb. 2023.

[19] V. Persson and J. Nouri, "A systematic review of second language learning with mobile technologies," *Int. J. Emerg. Technol. Learn.*, vol. 13, no. 2, pp. 188–210, 2018.

[20] E. L. Deci and R. M. Ryan, "The general causality orientations scale: Self-determination in personality," *J. Res. Personality*, vol. 19, no. 2, pp. 109–134, Jun. 1985.

[21] E. L. Deci and R. M. Ryan, "The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior," *Psychol. Inquiry*, vol. 11, no. 4, pp. 227–268, Nov. 2000.

[22] R. M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being," *Amer. Psycholog.*, vol. 55, no. 1, pp. 68–78, Jan. 2000.

[23] M. S. McEown and W. L. Q. Oga-Baldwin, "Self-determination for all language learners: New applications for formal language education," *System*, vol. 86, Nov. 2019, Art. no. 102124.

[24] P. Benson, *Teaching and Researching: Autonomy in Language Learning (Applied Linguistics in Action)*, 2nd ed. Evanston, IL, USA: Routledge, Nov. 2013.

[25] R. Godwin-Jones, "Emerging technologies autonomous language learning," *Lang. Learn. Technol.*, vol. 15, no. 3, pp. 4–11, Oct. 2011.

[26] Y. K. Dwivedi, D. L. Hughes, C. Coombs, I. Constantiou, Y. Duan, J. S. Edwards, B. Gupta, B. Lal, S. Misra, P. Prashant, R. Raman, N. P. Rana, S. K. Sharma, and N. Upadhyay, "Impact of COVID-19 pandemic on information management research and practice: Transforming education, work and life," *Int. J. Inf. Manage.*, vol. 55, Dec. 2020, Art. no. 102211.

[27] R. Huang, D. Liu, A. Tlili, and J. Yang, *Handbook on Facilitating Flexible Learning During Educational Disruption: The Chinese Experience in Maintaining Undisrupted Learning in COVID-19 Outbreak*. Beijing, China: Smart Learning Institute of Beijing Normal Univ., Mar. 2020.

[28] C. Chen, "Using scaffolding materials to facilitate autonomous online Chinese as a foreign language learning: A study during the COVID-19 pandemic," *SAGE Open*, vol. 11, no. 3, pp. 1–12, Aug. 2021.

[29] L. K. Visperas and Y. Chodpathumwan, "Time-series database benchmarking framework for power measurement data," in *Proc. Res., Invention, Innov. Congr., Innov. Elect. Electron. (RI2C)*, Sep. 2021, pp. 25–30.

[30] E. Klitzke. (Jul. 2016). *Why UBER Engineering Switched From Postgres to MySQL*. [Online]. Available: https://eng.uber.com/mysql-migration/

[31] S. S. Skiena, *The Data Science Design Manual*, 1st ed. Cham, Switzerland: Springer, Aug. 2017.

[32] R. P. Biuk-Aghai, R. C. K. Chan, Y.-W. Si, and S. Fong, "Visualizing recent changes in Wikipedia," *Sci. China Inf. Sci.*, vol. 56, no. 5, pp. 1–15, May 2013.

[33] S. Afonso and F. Almeida, "Fancier: A unified framework for Java, C, and OpenCL integration," *IEEE Access*, vol. 9, pp. 164570–164588, 2021.

[34] A. Wu, Y. Wang, X. Shu, D. Moritz, W. Cui, H. Zhang, D. Zhang, and H. Qu, "AI4 VIS: Survey on artificial intelligence approaches for data visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 12, pp. 5049–5070, Dec. 2022.

[35] A. Sahay, *Bus. Analytics: A Data-Driven Decision-Making Approach for Bus., vol. II*, 1st ed. Bus. Expert Press, Nov. 2019.

[36] A. B. Palacio, *The Art of Data-Driven Business: Transform Your Organization Into a Data-Driven One With the Power of Python Machine Learning*, 1st ed. Birmingham, U.K.: Packt Publishing, Dec. 2022.

[37] A. Tiwari, N. Sengar, and V. Yadav, "Next word prediction using deep learning," in *Proc. IEEE Global Conf. Comput., Power Commun. Technol. (GlobConPT)*, Sep. 2022, pp. 1–6.

[38] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *Social Netw. Comput. Sci.*, vol. 2, no. 6, pp. 1–20, Nov. 2021.

[39] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. (2018). *Improving Language Understanding by Generative Pre-Training*. OpenAI. [Online]. Available: https://api.semanticscholar.org/CorpusID:49313245

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[41] R. Thoppilan et al., "LaMDA: Language models for dialog applications," 2022. *arXiv:2201.08239*.

[42] C.-P.-E. M. T. Laboratory. (Feb. 2018). *Diz lá! Mobile Application*. [Online]. Available: https://cpelab.mpu.edu.mo/dizla/

[43] J. C. de Azeredo, *Dicionário Houaiss de Conjugação de Verbos*, 1st ed. São Paulo, Brazil: Publifolha, Jan. 2012.

[44] J. Han, J. Pei, and H. Tong, *Data Mining: Concepts and Techniques*, 4th ed. San Mateo, CA, USA: Morgan Kaufmann, Oct. 2022.

[45] L. Manelli and G. Zambon, *Beginning Jakarta EE Web Development: Using JSP, JSF, MySQL, Apache Tomcat for Building Java Web Appl.*, 3rd ed. New York, NY, USA: Apress, Aug. 2020.

[46] IP API. (2023). *Ip Geolocation API*. [Online]. Available: https://demo.ip-api.com/

[47] N. Algiryage, G. Dias, and S. Jayasena, "Distinguishing real Web crawlers from fakes: Googlebot example," in *Proc. Moratuwa Eng. Res. Conf. (MERCon)*, May 2018, pp. 13–18.

[48] A. Gorelik, *The Enterprise Big Data Lake: Delivering Promise Big Data Data Science*, 1st ed. Sebastopol, CA, USA: O'Reilly Media, Mar. 2019.

[49] Oracle. (2023). *MySQL Reference Manual*. [Online]. Available: https://dev.mysql.com/doc/refman/8.0/en/

[50] Timescale. (May 2023). *Get Started With Timescale*. [Online]. Available: https://docs.timescale.com/about/latest/

[51] K. J. Cios, R. W. Swiniarski, W. Pedrycz, and L. A. Kurgan, *Data Mining: A Knowledge Discovery Approach*, 1st ed. Cham, Switzerland: Springer, Dec. 2007.

[52] M. Amundsen, *RESTful Web API Patterns and Practices Cookbook: Connecting and Orchestrating Microservices and Distributed Data*, 1st ed. Sebastopol, CA, USA: O'Reilly Media, Nov. 2022.

[53] A. Géron, *Hands-On Machine Learning With Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, Oct. 2019.

[54] R. P. Biuk-Aghai, W. T. Kou, and S. Fong, "Big data analytics for transportation: Problems and prospects for its application in China," in *Proc. IEEE Region 10 Symp. (TENSYMP)*, May 2016, pp. 173–178.

[55] D. Perino, M. Varvello, and C. Soriente, "ProxyTorrent: Untangling the free HTTP(S) proxy ecosystem," in *Proc. World Wide Web Conf. World Wide Web (WWW)*, 2018, pp. 197–206.

[56] F. Laricchia. (May 2023). *Global Smartphone Market Share From Fourth Quarter 2009 to First Quarter 2023*. [Online]. Available: https://www.statista.com/statistics/271496/

[57] Needham. (Jan. 2023). *Smartphone Market Share*. [Online]. Available: https://www.idc.com/promo/smartphone-market-share/

[58] BBC. (Nov. 2022). *US Bans Sale of Huawei, ZTE Tech Amid Security Fears*. [Online]. Available: https://www.bbc.com/news/world-us-canada-63764450/

[59] UNNews. (May 2023). *Who Chief Declares End to COVID-19 as a Global Health Emergency*. [Online]. Available: https://news.un.org/en/story/2023/05/1136367/

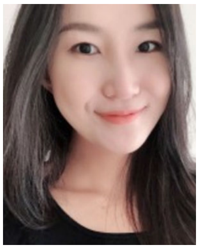
[60] C. Cunha and L. Cintra, *Nova Gramática do Português Contemporâneo*, 7th ed. Little Rock, Arkansas: Lexikon, May 2021.

[61] D. J. Skrien, *Object-Oriented Design Using Java*, 1st ed. McGraw-Hill Education, Jan. 2008.

[62] E. B. P. Raposo, M. F. B. do Nascimento, M. A. C. da Mota, L. Segura, and A. Mendes, *Gramática do Português*, vol. 1, 1st ed. Lisbon, Portugal: Fundação Calouste Gulbenkian, Jan. 2013.

[63] E. Elrom, *Integrating D3.Js With React: Learn to Bring Data Visualization to Life*, 1st ed. New York, NY, USA: Apress, Jun. 2021.

LAP MAN HOI (Member, IEEE) received the bachelor's degree in computer science from York University, Canada, and the master's degree in internet computing from the Queen Mary University of London. He is currently pursuing the Ph.D. degree in computer applied technology with the Faculty of Applied Sciences, Macao Polytechnic University (MPU). He was a researcher in the field of gaming and entertainment. He is currently a Researcher in the field of machine translation with the Faculty of Applied Sciences, MPU. His research interests include internet computing, data warehouse, data science, gaming, deep learning, machine translation, and voice recognition.



YUQI SUN received the master's and Ph.D. degrees from Faculdade de Letras, University of Lisbon, Portugal. She is currently a Lecturer with the Faculty of Applied Sciences, Macao Polytechnic University. Her research interests include applied linguistics, second language acquisition, pragmatics, systemic functional linguistics, and translation studies.

WEI KE (Member, IEEE) received the Ph.D. degree from the School of Computer Science and Engineering, Beihang University. He is currently a Professor with the Faculty of Applied Sciences, Macao Polytechnic University. His research interests include programming languages, image processing, computer graphics, tool support for object-oriented and component-based engineering and systems, the design and implementation of open platforms for applications of computer graphics, and pattern recognition, including programming tools, environments, and frameworks.

SIO KEI IM (Member, IEEE) received the degree in computer science and the master's degree in enterprise information systems from the King's College, University of London, U.K., in 1998 and 1999, respectively, and the Ph.D. degree in electronic engineering from the Queen Mary University of London (QMUL), U.K., in 2007. He gained the position of a Lecturer with the Computing Programme, Macao Polytechnic Institute (MPI), in 2001. In 2005, he became the Operations Manager of MPI-QMUL information systems research center jointly operated by MPI and QMUL, where he carried out signal processing work. He was promoted to a Professor with the Macao Polytechnic Institute, in 2015. He was a Visiting Scholar with the School of Engineering, University of California at Los Angeles (UCLA), and an Honorary Professor with The Open University of Hong Kong.