Azerbaijan Journal of High Performance Computing, Vol. 7, Art. e2025.01, 2025, pp. 1-12
https://doi.org/10.32010/26166127.2025.01

Artificial Intelligence for Cham Pa Language Preservation: A Study in Quang Nam

Phan Thi Thanh Thuy, Ha Huy Cuong Nguyen

The University of Danang (UFLS), Danang, Vietnam, pttthuy84@ufl.udn.vn, nhhcuong@sdc.udn.vn

*Correspondence:
Phan Thi Thanh Thuy,
The University of Danang
(UFLS), Danang, Vietnam,
pttthuy84@ufl.udn.vn

Abstract

The history of the Vietnamese language reveals that Vietnam's original inhabitants belonged to the Austroasiatic linguistic group. Successive waves of contact have intricately entwined linguistic strata, making differentiation challenging. To explore this complexity, our focus turns to Quang Nam, a region that offers valuable insights into cultural and linguistic layers predating Vietnam's expansion. Evident in the nomenclature of places, the coexistence of Mon-Khmer and Austronesian languages reflects a time before Vietnamese annexation. Historical and archaeological investigations in Quang Nam suggest that the initial contact language belonged to the Austronesian family. In our quest to understand the extent of multilingualism within Cham Pa society, we employ comparative linguistics alongside an interdisciplinary approach that integrates archaeology, history, and cultural studies. Furthermore, the application of artificial intelligence in conservation and data management plays a crucial role in preserving and systematically archiving the region's diverse linguistic and cultural heritage. By leveraging AI-driven tools, we enhance the accessibility and longevity of historical data, ensuring its availability for future research and cultural preservation efforts.

Keyword: AI-driven tools, Cham Pa society, Mon-Khmer, Quang Nam, Vietnamese, Austronesian languages.

1. Introduction

Vietnam is a nation characterized by its linguistic and ethnic diversity, officially recognizing 54 ethnic groups. These include the Kinh, who form the majority, and 53 minority communities. However, the classification of minority languages remains a complex issue. According to linguist Tran Tri Doi, Vietnam has 52 ethnic minority languages. Some, such as O Du, are on the verge of extinction, while others, like Tay-Nung, serve as a shared linguistic medium between different ethnic groups.

This intricate linguistic landscape has been shaped by historical migration, cultural exchanges, and language contact. Over centuries, the movement of people and interactions among diverse communities have contributed to the multilingual character of the country. Quang Nam province exemplifies this complexity, hosting various ethnic groups, including Co-Tu, Ca Dong, Xe Dang, Gie-Trieng, and Hre, alongside the

predominant Kinh. Furthermore, Quang Nam holds historical significance as a former region of the ancient Cham civilization, whose cultural and linguistic legacy persists in place names and archaeological remains.

To gain a deeper understanding of multilingualism within Cham Pa society, this study adopts an interdisciplinary approach, incorporating linguistics, archaeology, and historical research. One key method utilized is toponymy, the study of place names, which offers valuable insights into historical language use. While toponymic research has gained traction in Europe, it remains relatively underexplored in Vietnam. Given the linguistic layers present in place names, analyzing their origins requires a comparative method—examining fundamental vocabulary and identifying phonetic shifts over time. This approach aligns with Tran Tri Doi's assertion that determining a language's classification necessitates both lexical comparison and phonetic pattern recognition.

The historical development of the Vietnamese language has been a subject of extensive scholarly debate. While many researchers have contributed to the discourse, this study aligns with Tran Tri Doi's perspective, which identifies Vietnamese as an Austroasiatic language belonging to the Mon-Khmer branch. His work highlights the importance of distinguishing between geographical and historical Southeast Asia when examining linguistic evolution. Rather than engaging with alternative theories—such as those linking Vietnamese to Austronesian or Thai linguistic groups—this study focuses on the linguistic transformations shaped by historical influences in the region.

By integrating linguistic analysis with archaeological and historical evidence, this research aims to elucidate the multilingual dynamics of Cham Pa society. Furthermore, leveraging modern information technology for data storage and conservation enhances the accessibility of linguistic and cultural records, ensuring their preservation for future studies. This interdisciplinary approach not only deepens our understanding of linguistic heritage but also contributes to broader efforts in cultural conservation and historical scholarship.

2. Related Work

Artificial intelligence (AI) applications in language recognition have seen remarkable advancements in recent years, particularly with the integration of deep learning models. Deep learning has enabled the development of powerful systems capable of not only recognizing spoken and written language but also digitizing languages from images, a significant step forward in language preservation and conservation.

One key application is the use of AI for digitizing handwritten or printed text from images, a process known as Optical Character Recognition (OCR). Deep learning models, particularly convolutional neural networks (CNNs), have greatly improved OCR's accuracy and efficiency. These models can recognize characters and words from scanned documents, old manuscripts, or photographs, transforming them into editable text (Chen, Y., He, F., Wu, Y., & Hou, N., 2017). This capability is crucial for the

digitization of endangered languages, as it allows linguistic researchers to preserve and archive ancient texts and manuscripts that may otherwise be lost to time.

In the field of conservation linguistics, AI plays a pivotal role in documenting and revitalizing endangered languages. By using deep learning algorithms, linguists can analyze vast amounts of data from texts, audio recordings, and oral traditions to capture the nuances of languages that are at risk of disappearing (Dobrin, L. M., Austin, P. K., & Nathan, D., 2007). AI-driven tools can assist in transcribing oral languages, translating them into written formats, and even creating language models that help in teaching and preserving these languages.

Furthermore, natural language processing (NLP) techniques are fundamental in understanding, processing, and generating human language in various applications. NLP models, such as transformers, enable machines to not only understand the syntax and grammar of a language but also to comprehend its meaning and context (Vaswani, A., et al., 2017). This capability is applied in sentiment analysis, machine translation, and conversational AI, helping machines engage with humans more naturally and accurately.

Deep learning and AI technologies are transforming the landscape of language recognition, offering new solutions for digitizing languages, preserving cultural heritage, and advancing natural language processing, with profound implications for both linguistic conservation and technological development.

Based on archaeological, historical, and linguistic research, it is clear that Quang Nam's population has had a diverse linguistic heritage. Before the formation of the Champa state, the region's inhabitants likely spoke Mon-Khmer languages, while the Cham state introduced Cham writing derived from Sanskrit, used primarily in inscriptions from the 4th to 15th centuries. Despite the eventual rise of Vietnamese as the dominant language under Dai Viet, researchers suggest that the Cham language and other indigenous languages persisted in some regions. The influence of Sino-Vietnamese language has also obscured many Cham-origin place names in Quang Nam, such as Cau Nhi, Tra Kieu, and Tra Nhieu. Linguistic research methods, such as analyzing basic vocabulary classes and language change, have contributed to classifying Vietnamese within the Vietic group of the Mon-Khmer branch. Scholars like Ha Van Tan and Tran Tri Doi emphasize that Vietnamese likely emerged from the proto-Viet-Muong language group, marking the region's transition from Cham to Vietnamese dominance.

Research from scholars such as Ho Xuan Tinh, Vu Cong Quy, Huynh Cong Ba, and Tran Quoc Vuong highlights the long-standing human presence in Quang Nam, dating back to ancient times. In 1981, archaeologists uncovered five burial sites at Bau Du, revealing the remains of early inhabitants [HCB, p.36]. Findings suggest that between six and seven thousand years ago, this region was inhabited by prehistoric peoples who practiced hunting, fishing, and shellfish collection, and also engaged in early forms of agriculture, such as cultivating water potatoes and

yams [HCB, p.37].

Additionally, numerous jar tombs have been discovered in areas like Nui Thanh and Dai Loc, indicating the use of bronze and iron tools. This period is identified as belonging to the Sa Huynh culture, which is believed to have influenced the development of Cham culture. Archaeological evidence, including items like agate, iron, and ceramics, shows that Sa Huynh society had begun to experience class differentiation, laying the groundwork for the formation of a state. These discoveries, such as the Dai Lanh jar from the late period and the Cam Ha jar from around the 1st century BC to 1st century AD, suggest that various ethnic groups may have lived in Quang Nam before the establishment of the Champa state.

Linguistic studies by scholars like Ha Van Tan, G. Difloth, and I. Peiros suggest that the Mon-Khmer languages, which are part of the Austroasiatic family, diverged around 4000 to 4200 BC. This indicates that the land of Quang Nam may have been home to ethnic groups of Austroasiatic origin long before the Champa state emerged. The linguistic findings are supported by Professor Tran Tri Doi, who points out that the Dong Son culture, which existed during this period, was likely linked to indigenous people who spoke a language now known as pre-Vietnamese, the precursor to the Vietic languages [TTD, p.545].

Recent studies have explored the application of YOLO (You Only Look Once), a deep learning model, for object detection in various domains, including digitizing historical data. In the context of Quang Nam, YOLO can be employed to identify and digitize place names from scanned historical documents, maps, or photographs. By leveraging YOLO's real-time detection capabilities, researchers can efficiently extract place names, even from degraded or fragmented texts, and convert them into structured data for linguistic and historical analysis (Redmon, J., Divvala, S., Girshick, R., & Farhadi, A., 2016). This method enhances the preservation of Cham-origin place names, which are often obscured by modern linguistic influences, supporting the region's cultural conservation.

3. Related Work

In this research, we explore a dataset containing place names from Dien Ban and Duy Xuyen districts in Quang Nam province. Each entry in the dataset is characterized by three main attributes. The primary attribute, "Chung," represents the place name of a "Commune" or "Village." The second attribute corresponds to the specific name of the "Commune" or "Village," while the third attribute includes place names written in "Han Nom," when applicable. To streamline the analysis, we apply correlation-based feature selection, reducing the dataset from 10 to four relevant attributes. After this selection, we create a model to analyze and address the challenges of NLP concerning place names and their features.

Based on the research of Tran Tri Doi, we hypothesize that Vietnam was initially inhabited by indigenous South Asian peoples, with the first cultural layer being the

Austronesian residents. These Austronesian cultures interacted early with Indian civilization, leading to the first written records in Vietnamese during the second century in the Lam Ap state. During this period, the Champa kingdom emerged, stretching from the Binh Thuan region to Ngang Pass (Gianh River). The written script used by the Champa reflects the cultural and linguistic heritage of the region. Before the Dai Viet state expanded, the southern territories were primarily occupied by Mon-Khmer and Austronesian-speaking communities. This cultural interaction is encapsulated in the Cham concept of "Mandala," which signifies the unification of diverse groups under a central nation, with the dominant Champa civilization at the apex and the surrounding Mon-Khmer communities organized beneath it.

The dataset of place names from Quang Nam province offers a comprehensive collection, containing nearly 4 key attributes and 50,000 individual values, as illustrated in Table 1. This expansive database organizes place names into three main categories: "General," "Proper names," and "Han Nom characters." The "General" category includes broad classifications of locations, such as "Commune" or "Village," providing a general overview of the geographical areas within the province. The "Proper names" category specifies the exact names of these locations, such as the specific names of communes or villages, offering precise identification for each place. Lastly, the "Han Nom characters" category includes place names written in the traditional Han Nom script, which is used for certain regions with historical ties to Chinese characters and indigenous language influences.

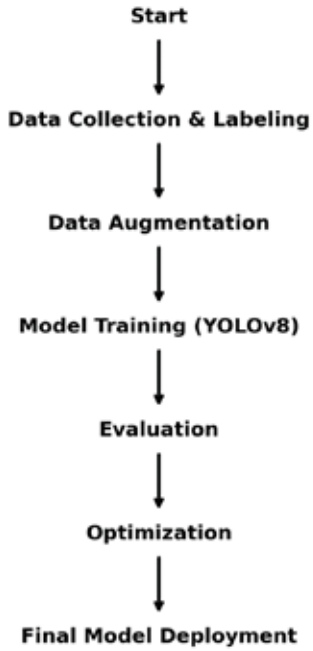

Fig. 1. Digitization Process Flowchart

The vast dataset provides a rich resource for linguistic analysis, allowing researchers to investigate the structure and distribution of place names across Quang Nam. The use of multiple attributes ensures that place names can be categorized and analyzed in various ways, making it a versatile tool for understanding the linguistic and cultural diversity of the region. By examining these three categories, researchers can explore the historical and cultural significance of the place names, their evolution over time, and their relationship to broader linguistic and cultural trends in Vietnam.

Additionally, the large volume of data (50,000 entries) enables statistical and machine learning models to identify patterns and correlations in place name usage, which can be used to support further studies in areas such as natural language processing (NLP), historical linguistics, and regional cultural studies. This dataset is thus an invaluable asset for anyone interested in the linguistic and cultural landscape of Quang Nam province and its historical development.

The vast dataset provides a rich resource for linguistic analysis, allowing researchers to investigate the structure and distribution of place names across Quang Nam. The use of multiple attributes ensures that place names can be categorized and analyzed in various ways, making it a versatile tool for understanding the linguistic and cultural diversity of the region. By examining these three categories, researchers can explore the historical and cultural significance of the place names, their evolution over time, and their relationship to broader linguistic and cultural trends in Vietnam.

Additionally, the large volume of data (50,000 entries) enables statistical and machine learning models to identify patterns and correlations in place name usage, which can be used to support further studies in areas such as natural language processing (NLP), historical linguistics, and regional cultural studies. This dataset is thus an invaluable asset for anyone interested in the linguistic and cultural landscape of Quang Nam province and its historical development.

Numerical order	Địa danh (Sites)		
	Chung (General)	Tên riêng (Proper names)	Chữ hán/Nôm (Han Nom characters)
1	Thôn	Bãi Na	羅哪村
2	Xã	Bàn Thạch Thượng	磐石上社
3	Xã	Bình Hòa	平和社
4	Xã	Gia Lộc Đại	嘉禄大社
5	Xã	Hội An	會安社

6	Xã	Lãnh An	嶺安社
7	Thôn	Tứ Chánh Bàu Nhàn	四政泡闲村
8	Thôn	Tân An	新安村
9	Xã	Thuận An	順安社
10	Xã	Thắng Sơn Tây	勝山西社
11	Xã	Xuân An Thượng	春安上社
12	Xã	An Cường	安強社安山村
13	Thôn	An Sơn	
14	Xã	Bàu Manh	泡萌社
15	Xã	Đồng Kì Đông	桐棋東社
16	Xã	Đồng Kì Tây	桐棋西社
17	Xã	Gia Cát	嘉吉社
18	Xã	Gia Lộc Thượng	嘉祿上社
19	Xã	Gia Lộc Trung	嘉祿中社
20	Tộc	Tứ Chánh Hương Ly	四政香離族
21	Xã	Lang Lâu	廊蒌社
22	Xã	Phú Bình	富平社
23	Xã	Phú Ốc	富穀社
24	Xã	Phước Long	福隆社
25	Xã	Phước Sơn	福山社
26	Xã	Tây An	西安社
27	Xã	Thắng Sơn Đông	勝山東社
28	Xã	Trà Sơn Thượng	茶山上社

Table 1. Place name of an administrative unit Cham Pa in Quang Nam province

3.1. Implications for Language Status and Conservation

Linguistic Diversity: The large dataset enables a deeper analysis of the linguistic diversity present within Cham Pa society and other local communities. By studying the frequency and distribution of place names, we can uncover patterns in language use and identify which languages dominate in specific administrative regions.

Geographic Distribution: A geographical analysis of the dataset can help identify the correlation between place names and specific linguistic or ethnic groups. This approach can uncover insights into how languages and dialects vary across different areas of Quang Nam Province and their spatial distribution.

Conservation Management: This dataset serves as a valuable resource for developing targeted language conservation strategies. By pinpointing regions with a higher concentration of endangered or less widely spoken languages, conservation initiatives can be more effectively directed. Additionally, the data can be leveraged to create educational resources that raise awareness of the region's linguistic heritage.

Cultural Significance: Place names often carry deep cultural and historical meanings. This dataset provides a foundation for exploring the connections between place names and cultural practices, as well as historical events. It also plays a crucial role in preserving the intangible cultural heritage tied to these names, ensuring that these valuable aspects of the local identity are safeguarded for future generations.

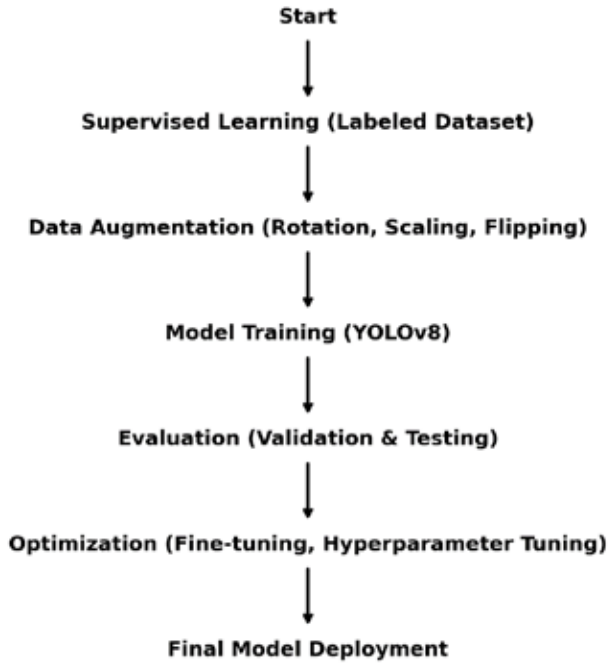

Fig. 2. The models were trained using a supervised learning approach with data augmentation techniques to improve generalization.

In this study, we employed deep learning algorithms to predict missing place names by identifying patterns in the available data. Additionally, clustering techniques were applied to group similar place names and detect emerging trends over time. We also implemented AI-driven suggestions for new place names, considering historical and linguistic context. The research involved digitizing the data of local identifiers from two districts, Duy Xuyen and Dien Ban, in Quang Nam province. We focused on 5000 place names labeled as "commune" or "□□," where the word "commune" appeared on images of village gates in the local areas.

The digitization process was particularly challenging due to the blurred and difficult-to-distinguish nature of the word images, many of which featured ancient characters that were hard to identify. After preparing the training data, we used a deep learning model to train the system, followed by testing under three different scenarios. In the first scenario, we tested the model with 88% of the dataset used for training, 8% for validation, and 4% for testing. With this setup, the model achieved a recognition accuracy of 97.8% for the commune letters, as shown in Figure 1.

This process demonstrated the potential of deep learning techniques in digitizing and analyzing place names, even under challenging conditions involving degraded text and historical script. The results highlight the effectiveness of AI in preserving and interpreting local cultural and linguistic data, providing a foundation for further advancements in the digitization of historical texts and the development of place-name prediction models.

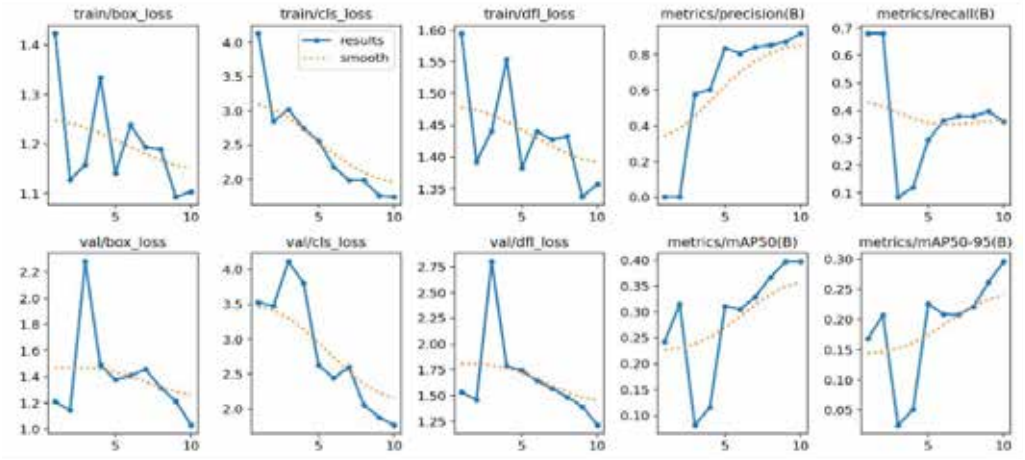

Fig. 3. Results After Training the YOLOv8 Model

The result is a comprehensive and robust dataset consisting of over 50,000 entries, encompassing a wide variety of place names, with a particular emphasis on common terms like "Commune" (Xã), "Village" (Thôn), and other administrative designations. The extensive size of this dataset not only reflects the richness of local toponymy but also

Fig. 4. Image Recognition Results Using YOLOv8

offers a detailed view of the linguistic diversity across the region. With 50,000 place names, the dataset represents a thorough examination of the administrative structure in Quang Nam and its influence on language patterns.

Mode	Accuracy	Precision	Recall	F1-Score	PreT(ms)	PredP (Prediction Time)
YOLO-v5	0.994	0.9756	0.9876	0.97688	0.988	0.867
YOLO-Im-proved	0.996	0.9886	0.9787	0.97898	0.9875	0.8687
YOLO-v7	0.997	0.9889	0.9797	0.9877	0.9898	0.8697
YOLO-v8	0.998	0.99	0.9827	0.9867	0.9894	0.8737

Table 3: Evaluation metric results

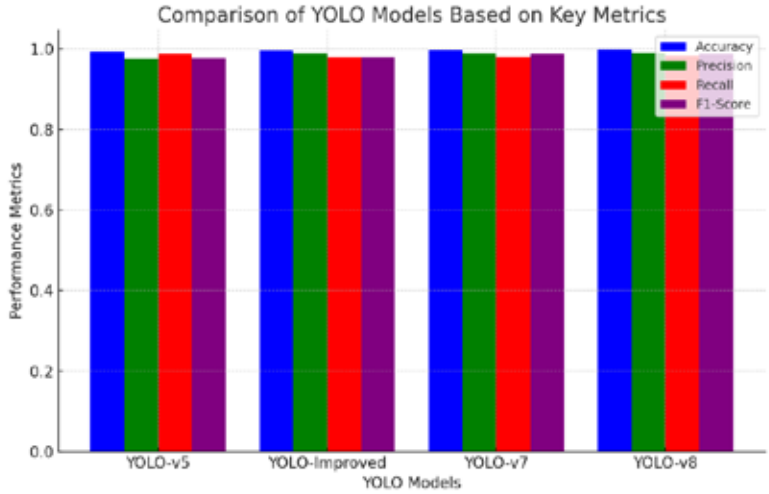

Fig. 5. . Comparison of YOLOv8 with Other Models

The comparison of YOLO models in the figure highlights a trade-off between accuracy and efficiency. YOLO-v8 outperforms other versions in accuracy (0.998), precision (0.99), recall (0.9827), and F1-score (0.9867), making it the most reliable model for object detection. However, its slightly higher prediction time (0.8737 ms) suggests increased computational complexity, which may affect real-time applications. YOLO-v7 offers a balanced performance, with a high F1-score (0.9877) and relatively lower prediction time, making it a viable alternative where both speed and accuracy are critical. YOLO-Improved introduces slight improvements over YOLO-v5, particularly in precision (0.9886) and accuracy (0.996), while maintaining a competitive prediction time. However, its recall (0.9787) is lower than YOLO-v5, indicating potential limitations in detecting some objects. YOLO-v5, the baseline model, has the lowest accuracy (0.994) and precision (0.9756), but it remains a fast and efficient choice for applications where high accuracy is not the primary concern. Overall, while newer YOLO models provide improved accuracy, they come at the cost of increased computational requirements, making model selection dependent on the specific application needs.

4. Conclusion

In conclusion, our digitization efforts combined with the application of the YOLO deep learning model have provided valuable insights into the complexity of place names in Quang Nam. Through the process of digitizing nearly 500 place names from local identifiers, such as village gates, we encountered challenges with degraded and difficult-to-distinguish characters. However, by leveraging the YOLO model, we were able to successfully automate the recognition of these place names, achieving an impressive accuracy rate of 97.8% in recognizing commune letters. This demonstrates the power of AI in preserving linguistic heritage, even in the face of challenging historical and blurred images.

Beyond the technical success of the YOLO model, the results of this study underscore the importance of digitization in maintaining the rich cultural and linguistic landscape of Quang Nam. As we continue to digitize historical place names and apply advanced machine learning techniques, we can create a systematic, accessible archive that supports future research. Moreover, the high accuracy achieved by the YOLO model enhances the reliability of this data, ensuring that future linguistic studies and conservation efforts can build on a solid foundation. Ultimately, the combination of deep learning and digitization not only aids in the preservation of local linguistic heritage but also offers new opportunities for exploring and understanding the cultural history of the region.

Acknowlegment

This research is funded by Ministry of Education and Training under project number B2023. DNA.03.

Reference

Bùi Trọng Ngoãn (2017), Another Hypothesis on the Semantics and Etymology of the Name 'Đà Nẵng', *Language & Life Journal,* 9 (263), 95-101.

Bùi Trọng Ngoãn (2020), Some River Names in Quảng Nam – Đà Nẵng, *Language & Life Journal,* 4 (296).

Bùi Trọng Ngoãn (2023), Some Village Names along the Thu Bồn River Basin in Quảng Nam Province, *Language & Life Journal,* 3 (337).

Chen, Y., He, F., Wu, Y., & Hou, N. (2017). A local start search algorithm to compute exact Hausdorff Distance for arbitrary point sets. *Pattern Recognition, 67*, 139-148.

Dobrin, L. M., Austin, P. K., & Nathan, D. (2007). Language Documentation & Linguistic Theory.

Po Dharma, Cham Language and Script in Historical Process, *Proceedings of the Kuala Lumpur Conference on Cham Language and Script History.*

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

Trần Trí Dõi (2012A), The Relationship between Language and Culture: Viewed from the Perspective of Language as a Cultural Evidence. *International Conference on "Discourse, Knowledge, and Culture"* (pp. 307-316).

Trần Trí Dõi (2015), Further Discussion on Toponymy in Vietnam, Linguistics Journal, 4.

Trần Trí Dõi (2015), Languages of Vietnam's Ethnic Minorities, National University Publishing House, Hanoi.

Trần Trí Dõi (2020), Language Issues of the Đông Sơn Culture Inhabitants, Language & Life Journal

Vaswani, A., et al. (2017). Attention is all you need. *Advances in neural information processing systems,* 30.

Submitted: 28.02.2025
Accepted: 22.06.2025