Addressing Regional Variation in Pidgin for Web Semantics

Peace B. Falola[1,*,†], Jesutomi Orija[2,†] and Olufade F. W. Onifade[1,†]

[1]Department of Computer Science, University of Ibadan, Ibadan, Nigeria
[2]Department of English, University of Ibadan, Ibadan, Nigeria

Abstract

This paper investigates the integration of Pidgin, a widely spoken language in West Africa, into web semantics to address linguistic inclusivity in digital spaces. Focusing on regional variations, the study examines syntactic, lexical, and semantic inconsistencies across dialects of Pidgin, emphasizing challenges in standardization and ontology mapping. Using a hybrid methodology combining corpus analysis and semantic modeling, the research identifies patterns in regional differences and proposes an adaptable framework for language processing. A corpus of Pidgin data was collected from diverse sources, including online forums, social media platforms, and news websites, using automated web-scraping techniques. The extracted data were analyzed for context-specific meanings and structural variations. A semantic annotation schema was developed to facilitate machine-readable interpretations, enabling integration into semantic web technologies. The findings highlight the importance of linguistic diversity in digital applications and propose tools for enabling localized Pidgin content retrieval and classification. Future work includes refining ontological mappings and expanding datasets to improve the scalability and robustness of Pidgin language processing systems.

Keywords

Pidgin, Semantic Web, Regional Variations, Information Retrieval

1.0 Introduction

Pidgin English (PE) has developed and has been used over the years as a result of changing historical and cultural trends, which have also had a major impact on the sociolinguistic elements that influence its use. Research indicates that in West Africa, PE has gradually improved cross-cultural and inter-ethnic communication [1,2,3]. In addition to encouraging inter-ethnic dialogue, Umana [4] contended that PE has made a substantial contribution to the region's cultural fusion. Dynamic sociolinguistic elements including socioeconomic classes, polyglotism, and differing cultural worldviews and identities also have an impact on the use of PE [5,6]. PE has changed over the past several years and is now spoken in a number of countries, including Equatorial Guinea, Ghana, Cameroon and Nigeria, according to Osoba [7] and Yakpo [8]. There are notable regional differences in PE, which have been attributed to the impact of regional languages and dialects.

Numerous pidgins originated from colonial interactions that necessitated communication between the dominated and dominant cultures. PE originated during the colonial era and developed from a simple commercial language to a more sophisticated and dynamic mode of communication that helped to reduce linguistic and ethnic boundaries in West Africa. Regional variations in PE are a result of the diverse linguistic structures of the indigenous cultures that are predominant in those locations. Because of the diversity of PE in Nigeria, pidgin has been extensively discussed in relation to the development of languages and other forms of communication. Over time, certain pidgins have stabilized, allowing them to be recognized as a language that may be passed down from generation to generation. Beyond

MSW-24: 4th International Workshop on Multilingual Semantic Web, December 10, 2024, co-located with 6th Knowledge Graph and Semantic Conference (KGSWC-2024), December 11-13, 2024, Paris, France.

*Corresponding author.
†These authors contributed equally.
peacefalola@gmail.com (P. B. Falola); orijajesutomi@gmail.com (J. Orija); olufadeo@gmail.com (O. F. W. Onifade)
0000-0001-8581-4123 (P. B. Falola); 0000-0003-4446-6518 (J. Orija); 0000-0003-4965-5430 (O. F. W. Onifade)

its role in colonial interaction, PE has evolved as a lingua franca among over 200 ethnic groups in Nigeria, which has had a considerable impact on its current state. Other pidgins include Haitian Creole, a pidgin that evolved from French and is spoken in Haiti, and Tok Pisin, which originated in Papa New Guinea. Because PE is a low-resource language, it must be integrated with digital frameworks such as Web Semantics. Web semantics, often referred to as the semantic web, is frequently seen as an expansion of the internet that organizes data so that computers can process, search, and interpret searches meaningfully. The incorporation of PE into online semantics will inevitably support the conservation of PE as a cultural asset and promote accessibility, inclusivity, and improved language representation. It will also support enhanced knowledge and data exchange, as well as promote PE and other pidgins research. The term "web semantics" describes the idea of giving meaning to data that is readily available on the internet to provide sufficiently precise information for both machine and human communication. Therefore, this paper addresses the integration of Pidgin English into web semantics and addresses the problem of regional varieties which pose a challenge to this endeavor.

2.0 Literature Review

Yakpo[8] discussed in their work how West African Pidgin ("Pidgin") is a cluster of related, mutually intelligible, restructured English with up to 140 million speakers in Nigeria, Cameroon, Ghana, Sierra Leone, Equatorial Guinea, and The Gambia. Spoken by just few thousand people two centuries ago, "modernization" and "shallow social entrenchment" have driven the transformation of Pidgin into a "super-central" world language. Demographic growth, migration, the expansion of West African cultural industries and economies, and people-to-people contacts are likely to expand Pidgin further. Already the largest language of West Africa, Pidgin may be spoken by 400 million people by 2100. The rise of Pidgin goes against the grain. World languages like English, French, Chinese, or Arabic mostly spread through colonization, elite engineering, and state intervention. The trajectory of Pidgin, therefore, holds great potential for exploring the dynamics of large-scale natural language evolution in the twenty-first century. Vats et al., [9] discussed in their work how knowledge graph (KG), a visual representation of text data as a semantic network, holds enormous promise for the development of more intelligent robots. It leads to significant potential solutions for many tasks like question answering, recommendation, and information retrieval. However, this area is confined to using English text only. Since low-resource languages are now being used in the world of AI, it is necessary to develop a semantic network for them as well. In this research work, the authors provide state-of-the-art techniques for automatic knowledge graph construction for the Hindi language, which is still unexplored in ontology. Constructing a knowledge graph faces several hurdles and obstacles in the linguistic domain, primarily when it deals with the Hindi language. With an emphasis on the Indian perspective, this research intends to introduce a novel approach 'HKG' for knowledge graph construction framework for Hindi. It also implements the LSTM model to evaluate the accuracy of newly constructed knowledge graphs and compute different evaluation metrics such as accuracy and F1-score. his knowledge graph evaluates the accuracy of 87.50 using Doc2Vec word embedding with a train-test split of 7:3.

Kaffee et al. [10] in their work discussed there is a lack of multilingual data to support applications in a large number of languages, especially for low-resource languages. The authors discussed how knowledge graphs (KG) could contribute to closing the gap of language support by providing easily accessible, machine-readable, multilingual linked data, which can be reused across applications. In their work, they provided an overview of work in the domain of multi-lingual KGs with a focus on low-resource languages. They reviewed the current state of multilingual KGs along with the different aspects that are crucial for creating KGs with language coverage in mind. Special consideration was given to challenges particular to low-resource languages in KGs. They further provided an overview of applications that yielded multilingual KG information as well as downstream applications reusing such multilingual data. Finally, they explored open problems regarding multilingual KGs with a focus on

low-resource languages.

Hampel, [11] discussed that student pidgin (SP) is an African youth language practice among Ghanaian students and graduates. It originated in cities along the Ghanaian coast, where most empirical research on SP has been conducted so far. Little is known about the use of SP in other regions of the country. Their work aimed to fill the research gap by comparing reported use and language attitudes of students in Ghana's two largest cities, the capital Accra and Kumasi, capital of the Ashanti region. The cities have a comparable number of inhabitants, but are located in different regions of Ghana. Over two hundred high school and university students answered a written questionnaire or participated in qualitative interviews. The results showed significant regional differences both in reported use and attitudes towards the youth language, which can be explained by the different language ecologies of Kumasi and Accra.

From the literature reviewed, it is evident that while existing studies address challenges faced by low-resource languages, they lack specific attention to the regional variations of Pidgin. Given the dynamic evolution of Pidgin, there is a pressing need for adaptable frameworks capable of constructing knowledge graphs and semantic networks that effectively capture its structural diversity and contextual usage. Although frameworks for other low-resource languages, such as Hindi, have been explored, no documented efforts have been made to develop similar frameworks tailored for Pidgin. Additionally, despite widespread acknowledgment of the scarcity of multilingual datasets for AI applications, corpus development and annotated datasets specific to Pidgin remain largely absent. Hampel (2020) highlighted regional differences in student Pidgin usage, underscoring the necessity of explicitly modeling these variations within web semantics. Finally, research in this area is still limited, presenting a significant gap that warrants focused attention from researchers.

3.0 Methodology

This study leveraged three variations of pidgin from 3 countries; Nigeria, Ghana and Cameroon. Figure 1 shows the framework for the study. Textual data were employed in this investigation. Textual information may be obtained from a number of sources, including blogs, academic materials, social media, news platforms (like BBC Pidgin), and online communities in the targeted areas. To determine important Pidgin terms and phrases as well as their regional connotations, these data were further annotated. For every Pidgin dialect (such as Ghanaian and Nigerian), a regional lexicon is produced, highlighting lexical, grammatical, and semantic variations. After additional processing and cleaning, non-pidgin and unnecessary texts were removed from the data. For examination, each text was further divided into individual words or phrases. They were then divided into groups according to the nations' regional identifiers. Using a regional lexicon, categorization, and regional lexicon matching, the regional variances were further categorized according to the region of usage. Contextual analysis helped to clarify ambiguous terms that have varied meanings depending on where they are used. By building nodes and establishing associations, the categorized data was further represented in a knowledge graph for use in web semantics. Lastly, accuracy was made on the precision of regional classification and ambiguity resolution. Figures 2, 3, 4, 5, and 6 provide a few instances of how to categorize the classes into different areas according to the circumstances.

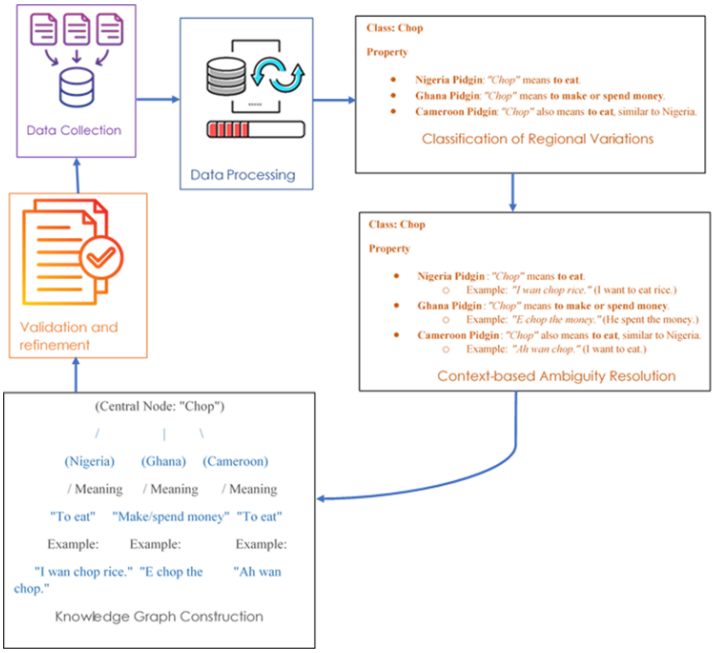

Figure 1: Framework of the Study

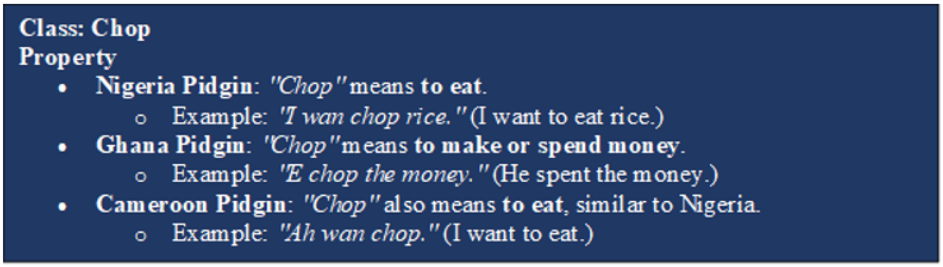

Figure 2: Example 1

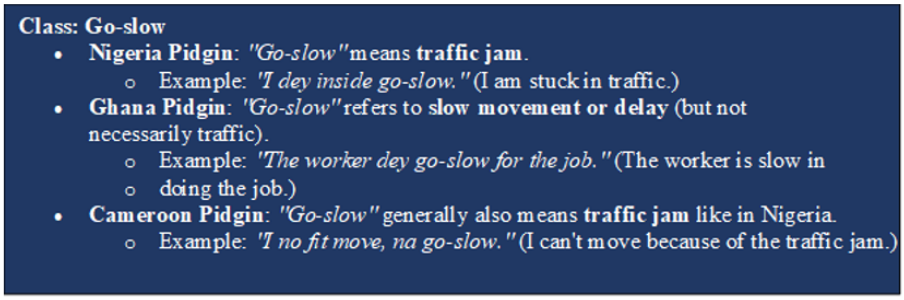

Figure 3: Example 2

Class: Dash

- **Nigeria Pidgin**: *"Dash"* means **to give something freely or as a gift**.
  - Example: *"I dash am the money."* (I gave him the money for free.)
- **Ghana Pidgin**: *"Dash"* means **to tip someone**.
  - Example: *"I go dash the taxi driver."* (I will give the taxi driver a tip.)
- **Cameroon Pidgin**: *"Dash"* means both **to give a gift** and **to bribe**.
  - Example: *"I dash the policeman small money."* (I gave the policeman a bribe.)

Figure 4: Example 3

Class: Wahala

- **Nigeria Pidgin**: *"Wahala"* means **trouble or problem**.
  - Example: *"No bring wahala come here."* (Don't bring trouble here.)
- **Ghana Pidgin**: *"Wahala"* also means **trouble**, but it's less commonly used compared to "Palava."
  - Example: *"E go cause wahala."* (He will cause trouble.)
- **Cameroon Pidgin**: *"Wahala"* has the same meaning of **trouble or problem**.
  - Example: *"This thing go bring wahala."* (This thing will cause trouble.)

Figure 5: Example 4

Class: Carry

- **Nigeria Pidgin**: *"Carry"* means **to transport or take something/someone**.
  - Example: *"I go carry you go school."* (I will take you to school.)
- **Ghana Pidgin**: *"Carry"* can also mean **to succeed in something**.
  - Example: *"E carry the competition."* (He succeeded in the competition.)
- **Cameroon Pidgin**: *"Carry"* often means **to pick up or hold**.
  - Example: *"Carry the bag well."* (Hold the bag properly.)

Figure 6: Example 5

Figures 7 to 11 show the representation of these examples in knowledge graphs which includes the nodes and edges which defines the relationship.

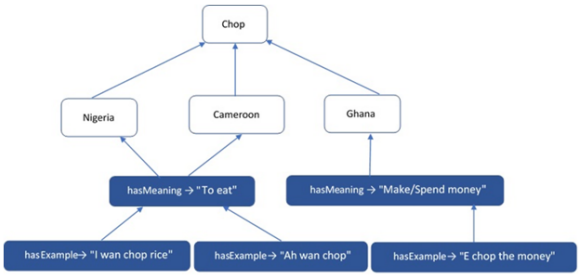

Figure 7: Knowledge Graph 1

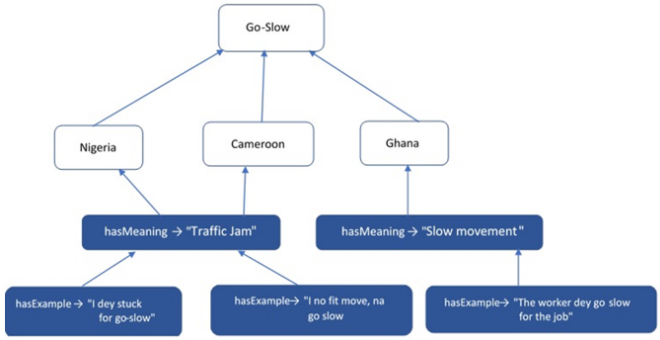

Figure 8: Knowledge Graph 2

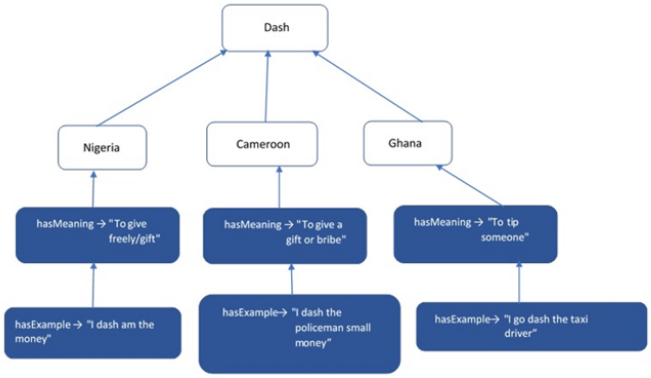

Figure 9: Knowledge Graph 3

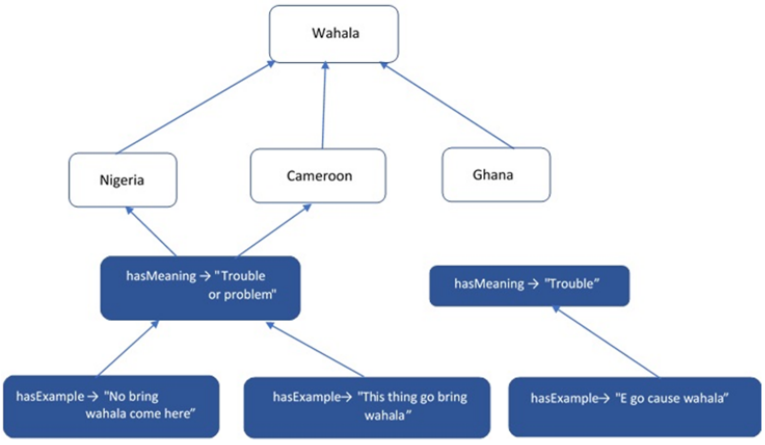

Figure 10: Knowledge Graph 4

The methodology is effective in addressing the key challenge of regional variation in Pidgin. By combining classification, contextual analysis, and knowledge graph construction, this approach provides a structured, human-centered solution for integrating Pidgin into Web Semantics. The knowledge graph serves as a crucial component for structuring Pidgin data in a way that is both machine-understandable and capable of resolving regional ambiguities, which ultimately improves information retrieval and interoperability across digital platforms.

4.0 Future Directions and Conclusion

There are a number of important avenues that may be taken in the future to broaden the study's effect and reach in terms of incorporating Pidgin into Web Semantics. The first primary focus is on diversifying the data sources. Future study will expand the data collection to include more region-specific social media platforms, radio transcripts, and spoken encounters from casual settings, even if the current technique depends on web scraping from a small number of online sources. By capturing a greater range of Pidgin usage, this extension would improve the dataset and more properly reflect the informal and geographical variances. Furthermore, the integration of automatic transcription technologies with audio data will provide a more thorough comprehension of the phonetic subtleties of Pidgin in various places, hence enhancing the semantic representation.

Making a common Pidgin ontology is another important step. It will be more accurate and useful to create a structured framework for the knowledge graph that takes into account regional variances. Better machine comprehension would be made possible by this ontology, which would standardize linkages between Pidgin words, meanings, and situations across geographical boundaries. It would also be beneficial to do cross-linguistic research contrasting Pidgin with similar regional tongues like Yoruba, Twi, or French Creole. Comprehending the ways in which regional variants of Pidgin are influenced by local languages may improve the system's capacity to manage multilingual situations and expand the semantic web of Pidgin.

Furthermore, the integration of feedback mechanisms driven by users may guarantee the ongoing development and enhancement of the knowledge graph. Pidgin speakers might update the graph with new terms, add missing words, or draw attention to regional differences that might not have been fully represented. This would allow the graph to change as the language did. This strategy would also make the tool more inclusive and encourage participation from the Pidgin-speaking community.

Future research will also look into enhancing automation with AI-based solutions. Optimized for Pidgin, natural language processing (NLP) models like BERT or GPT might automate certain aspects of the categorization and disambiguation process, increasing productivity, especially with enormous data sets. In real-time applications like chatbots or virtual assistants, where precise Pidgin comprehension is crucial, AI integration would be very helpful. Also, the establishment of an open-access regional Pidgin corpus would be a significant addition to this discipline. A corpus like this might help with computational linguistics research in the future and make it possible to create more advanced Pidgin-specific language technology. A well-annotated and standardized corpus would be the basis for many applications, such as instructional aids, translation systems, and text analysis.

In conclusion, this work offers a systematic methodology that integrates manual categorization, contextual ambiguity resolution, and knowledge graph creation to address the geographical variance in Pidgin. The technique makes sure that the semantic subtleties of Pidgin are accurately reflected by accounting for regional variations in Pidgin use throughout Nigeria, Ghana, and Cameroon. Although the current strategy has mostly depended on human experience, in order to scale the solution, future paths may make use of AI technology and user-driven contributions. The foundation for more extensive Web Semantics applications, such as improved search engines and multilingual systems, is laid by the development of a knowledge graph for Pidgin. This work advances digital inclusion for Pidgin speakers by incorporating Pidgin into the semantic web, supporting the larger endeavor to make web technologies relevant and accessible to low-resource languages.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly for Grammar and spelling check. Further, the author(s) used chatGPT 4.0 to get a concise summary for the abstract and also to summarize the paper for future directions and conclusions. The author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] Khan, I. J., & Akter, S. (2021). Pidgin and creole: concept, origin and evolution. *British Journal of Arts and Humanities*, **3**(6), 164-170.

[2] Kaviti, L., Oladipo, R., & Ndung'u, M. (2016). African Adaptation Processes in English: A Comparative Analysis of Nigerian Pidgin English and Kenyan "Engsh". *International Journal for Innovation Education and Research*, **4**(6), 50-66.

[3] Oguji, E.A. and Onuoha, J.N. (2022). 'Roles of Nigerian English Varieties in National Integration', *African Journal of Educational Management, Teaching and Entrepreneurship Studies*, **7**(2), pp. 216-223.

[4] Umana, B. F. H. (2018). Nigerian Pidgin English in Cape Town: exploring speakers' attitudes and use in diaspora. *Faculty of Humanities, Linguistics*.

[5] Ofulue, C. I. (2012). Nigerian Pidgin and West African pidgins: A sociolinguistic perspective. *Legon Journal of the Humanities*, **1**, 1-42.

[6] Dewi, K. T. (2021). Language use: Code mixing, code switching, borrowing, pidginization, and creolization. *Yavana Bhasha: Journal of English Language Education*, **4**(1), 34-44.

[7] Osoba, J. B. (2021). Nigerian Pidgin as national language: Prospects and challenges. *Current Trends in Nigerian Pidgin English: A Sociolinguistic Perspective*, **117**, 299.

[8] Yakpo, K. (2024). West African Pidgin: World Language Against the Grain. *Africa Spectrum*, **59**(2), 180-203.

[9] Vats, P., Sharma, N., & Sharma, D. K. (2023). Hkg: A novel approach for low resource Indic languages to automatic knowledge graph construction. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

[10] Kaffee, L. A., Biswas, R., Keet, C. M., Vakaj, E. K., & de Melo, G. (2023). Multilingual Knowledge Graphs and Low-Resource Languages: A Review. *Transactions on Graph Data and Knowledge*, **Vol. 1**, Issue 1, Article No. 10, pp. 10:1-10:19.

[11] Hampel, E. (2020). Regional variation in Ghanaian Student Pidgin: Use and attitudes. *Sociolinguistic Studies*, **14**(3).