Pioneering Word Embeddings for the Aja Language: Development and Insights

Josué Frejus Godeme
Dartmouth College
Hanover, NH, USA
josue.f.godeme.26@dartmouth.edu

Laura McPherson
Department of Linguistics
Dartmouth College
laura.mcpherson@dartmouth.edu

Rolando Coto-Solano
Department of Linguistics
Dartmouth College
Rolando.A.Coto.Solano@dartmouth.edu

Abstract

Note: This is not a published paper but a report on the work that has been accomplished.

This study presents the inaugural application of FastText word embeddings to the Aja language, aimed at uncovering semantic relationships in this underrepresented linguistic area. We developed FastText embeddings. Our evaluation, encompassing WordSimilarity-353 comparisons, distance visualization between pairs like 'eshi' (water) and 'ezo' (fire), and nearest-neighbor analysis for key terms such as 'Aja', highlighted the model's proficiency and limitations in semantic interpretation. The findings indicate promising avenues for leveraging FastText in less-represented languages, though they also point to the necessity of dataset expansion. This work contributes to natural language processing research and emphasizes the significance of linguistic diversity in AI.

1 Introduction

Africa is home to more than 2000 languages encompassing one-third of the world's languages estimated at around 6000 (Anderson, 2012). Despite that rich tapestry of languages, Africa's linguistic diversity is mostly ignored in research on multilingualism (Bylund et al., 2023) and critical fields like Artificial Intelligence. There are a lot of reasons to explain this bias, one being the lack of written text directly in African languages given Africa's history of oral transmission (Ki-Zerbo, 1969; Sone, 2018). Aja is a low-resourced African language part of the Niger-Congo family and the Gbe dialect continuum. The Aja language is spoken mostly in Benin, Togo, Ghana and Nigeria and counts around 1.3 million speakers (Eberhard et al., 2023). Despite having more than a million speakers, the Aja language is absent of most NLP applications. This paper addresses the significant gap in computational tools, specifically language embeddings, for Aja, a low-resourced language with minimal online presence and digital resources. The absence of such embeddings hinders the development of effective NLP applications, including machine translation and speech recognition, thereby marginalizing the speakers of the language. This research aims to develop the inaugural language embeddings for Aja, effectively bridging a crucial gap in NLP resources. The contributions of this work are unifold: it establishes a foundational resource that can catalyze further NLP research and applications for the Aja language.

2 Methodology

This section explains the methodology employed in developing embeddings for the Aja language. Our approach encapsulates a comprehensive process involving data collection, preprocessing, algorithm selection, and evaluation.

Given the scarcity of digital resources for Aja, our primary source was the SIL-published book "Lire et écrire l'ajagbe" (Beavon-Ham and Ega, 2007). This book, adheres to the spelling rules developed by the Comité International de Suivi de l'Orthographe de la Langue Aja (CISOLA) in conjunction with the Commission Nationale de Linguistique AJA (CNLA). It provided a corpus of 1453 words. We manually digitized these words, ensuring a reliable base for our embeddings. In addition to that, we asked an Aja native speaker to write 10 more sentences giving us another 95 words.

A critical preprocessing step was the removal of accents from the Aja text. This decision was driven by the potential to broaden the applicability of our embeddings. We acknowledge the linguistic compromise this entails and future work will explore the impact of this decision.

We chose the FastText algorithm (Joulin et al., 2016) instead of alternatives like Word2Vec (Mikolov et al., 2013). This decision was driven by FastText's ability to handle out of vocabulary words allows for better representation of rare words,

which is particularly advantageous given the limited size of our dataset.

Our methodology involved developing embeddings purely from the Aja corpus.

The embeddings were evaluated using a set of key Aja words, assessing both qualitative and quantitative aspects. Throughout the research, ethical considerations, particularly in terms of data usage and cultural sensitivity, were paramount. The source material's public availability and adherence to linguistic standards set by authoritative bodies in the Aja linguistic community underlined the ethical integrity of our methodology.

3 Experiments and Results

In this section, we explore the effectiveness and characteristics of the word embeddings created using our FastText model. Our investigation includes three different types of experiments: analysis using the WordSimilarity-353 Test Collection, a visual examination of the embeddings of the words 'Fire' and 'Water', and an exploration of the closest word neighbors for a selected word.

3.1 Experiment 1: WordSimilarity-353 Test

To validate my embeddings, I used the WordSimilarity-353 Test Collection, which provides human-assigned similarity scores for specific word pairs. The WordSim-353 Dataset is a dataset that can be used in a Semantic Word Similarity Benchmark Task (Finkelstein et al., 2002). I focused on the pair 'King' and 'Queen' from this test. The human-annotated score in WordSimilarity-353 for this pair is 8.58 ((0 = words are totally unrelated, 10 = words are VERY closely related)). In Aja, king is written as "efio" and queen as "efio si". In contrast, our FastText model yielded a cosine similarity score of 0.645 for the pair 'efio' and 'efio si'. This score suggests a moderate level of similarity, indicating that our model's understanding of these word relationships is somewhat aligned with human perception.

3.2 Experiment 2: Visual examination of the embeddings of the words 'Fire' and 'Water'

In Aja, the word "Fire" can be translated as "ezo" while the word "Water" can be translated as "eshi". In the context of semantic analysis and linguistic studies, "water" and "fire" are typically used as classic examples of contrasting elements due to their opposing natural properties and symbolic meanings. Our embeddings should then display that contrast if visualized. To visualize and interpret these high-dimensional embeddings, we applied t-SNE (t-distributed Stochastic Neighbor Embedding) for dimensionality reduction (van der Maaten and Hinton, 2008), followed by creating a scatter plot with Matplotlib (Hunter, 2007) to effectively illustrate the semantic relationships between words in a 2D space. The figure below shows that visualization. This visual is confirmed by the distance

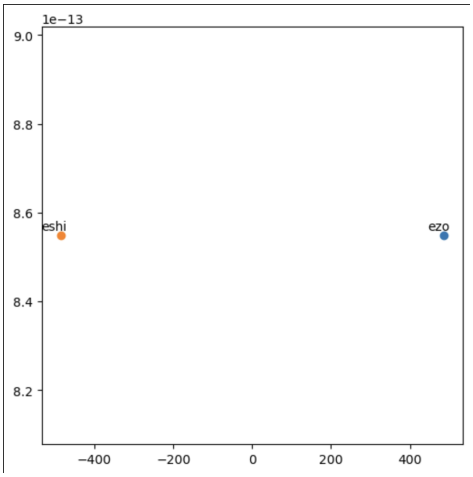

Figure 1: Representation of Water(Eshi) and Fire(Ezo)

between the two words when computing the cosine similarity.

In order to quantitatively assess the semantic relationship between words in our trained embeddings, we employed cosine similarity, a widely-used metric in natural language processing for determining the similarity between two vectors. Cosine similarity measures the cosine of the angle between two vectors in a multi-dimensional space, providing an efficient way to gauge the degree of similarity between word representations. For our analysis, we chose to focus on the words 'eshi' and 'ezo', which are of particular interest in our dataset (meaning water and fire respectively). We computed the cosine similarity between the word embeddings of 'eshi' and 'ezo' from our FastText model using Python's scipy.spatial.distance.cosine function. This analysis revealed a similarity score of -0.16, indicating that these words are highly dissimilar.

3.3 Experiment 3: Exploration of the closest word neighbors for a selected word

For this experiment, we chose the word "Aja" which simply designate the Aja Language like one could do with "French" and the "French Language". Using a python script, we displayed the 3 closest neighbors to the word "Aja". The table below displays the results.

Words	Similarity Score
ajagbe	0.34
sɔ	0.32
va	0.24

Table 1: Closest words to "Aja" according to our embeddings

The embeddings start right by displaying "adjagbe" as it means the "Adja language" which is the closest we have in our dataset to Aja. The 2 other words do not mean anything close to "Aja". Also, we need to notice that the similarity scores for "Ajagbe" for example is relatively low for such close words. Notably, the similarity scores, such as 0.34 for 'ajagbe', are relatively low. This observation underscores the potential limitations of our dataset's size, suggesting that a more extensive corpus might lead to higher confidence in the model's predictions and more accurate representations of semantic relationships. These findings point to the need for further research using larger, more diverse datasets to enhance the model's understanding of linguistic nuances.

4 Discussion

Our study leveraged the FastText model to build embeddings and analyze semantic relationships in for the Aja language. The experiments, particularly focusing on word pairs (WordSim-353 Dataset) and nearest neighbors, provided insight into the nuanced semantic structure captured by our embeddings. For instance, the close relationship between 'Aja' and 'ajagbe' underscores the model's capability in reflecting linguistic connections pertinent to the language.

A primary limitation of our study was the dataset's size, which likely influenced the model's performance and the lower confidence in predictions for certain word pairs. Additionally, unexpected findings, such as the low similarity scores for semantically close words, suggest areas for further refinement in our approach, possibly indicating the need for a more diverse or extensive corpus.

Conclusion

This study represents a pioneering effort in developing word embeddings for Aja using the FastText model. As a first attempt, it has laid the groundwork for understanding the semantic structure of Aja. The findings, particularly in identifying significant word relationships, offer initial insights into the language's complex semantic landscape. The implications of this research extend to the broader field of computational linguistics, especially in underscoring the potential of word embeddings in exploring less-represented languages. Recognizing that our approach is in its nascent stages and perfectible, future research should aim at expanding the corpus to encompass a more diverse linguistic spectrum of the Aja language. Refinement in embedding techniques and exploration of different model parameters could further enhance the quality of the embeddings. Practical applications of these embeddings in tasks like contextual analysis, spell checker and language translation could also be explored, providing a testbed for their real-world utility.

It is important to underscore that this is the first endeavor to create embeddings for Aja. As such, while the approach shows promise, it is inherently perfectible. We anticipate that subsequent research will build upon and refine this foundational work, progressively enhancing the linguistic and semantic understanding of Aja Language.

Ethics Statement

In accordance with the ACL Ethics Policy[1], our research on developing word embeddings for Aja upholds ethical standards. We ensured cultural sensitivity and fair representation in our linguistic analysis, adhered to data privacy norms. Our work contributes to linguistic diversity in natural language processing, promoting inclusivity for underrepresented languages. While we recognize the potential for the misuse of language technologies, we advocate for their responsible and beneficial use.

Acknowledgements

We extend our heartfelt thanks to The Neukom Institute at Dartmouth for their generous funding and

[1] https://www.aclweb.org/portal/content/acl-code-ethics

support of this research endeavor. Special gratitude is owed to Professor Laura McPherson, Associate Professor in the Linguistics program at Dartmouth College. Her extensive expertise in phonology, morphology, and fieldwork, particularly in African languages, were invaluable in guiding this project

Our sincere appreciation also goes to Professor Rolando Coto-Solano, a computational linguist. Professor Coto-Solano, with his robust background in computational linguistics, provided critical guidance in the computational aspects of this research.

Their combined insights and mentorship have been instrumental in the successful completion of this project.

References

Stephen R. Anderson. 2012. How many languages are there in the world? In Stephen Anderson, editor, *Languages: A Very Short Introduction*, page 0. Oxford University Press.

Virginia Beavon-Ham and Emile Ega. 2007. *Lire et écrire l'ajagbe: Guide pratique pour ceux qui savent lire le français*. SIL Bénin, Cotonou.

Emanuel Bylund, Zainab Khafif, and Robyn Berghoff. 2023. Linguistic and geographic diversity in research on second language acquisition and multilingualism: An analysis of selected journals. *Applied Linguistics*, page amad022.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2023. *Ethnologue: Languages of the World*, twenty-sixth edition. SIL International, Dallas, Texas.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

J. D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. ArXiv:1607.01759 [cs].

Joseph Ki-Zerbo. 1969. The Oral Tradition as a Source of African History. *Diogenes*, 17(67):110–124. Publisher: SAGE Publications Ltd.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. ArXiv:1301.3781 [cs].

Enongene Mirabeau Sone. 2018. African Oral Literature and the Humanities: Challenges and Prospects. *Humanities*, 7(2):30. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.