Lexical Diversity Across African Languages: A Contrastive Analysis of Igbo, Yoruba, Hausa, and Nigerian Pidgin

Submission Category: Poster

Abstract
Lexical diversity is a key measure of language complexity, reflecting variation in vocabulary use across discourse domains, sentence structures, and communicative functions. This study compares lexical diversity in Igbo, Yoruba, Hausa, and Nigerian Pidgin using spoken discourse transcriptions. The analysis applies Type-Token Ratio (TTR) and Measure of Textual Lexical Diversity (MTLD) to examine how linguistic structures affect vocabulary richness in agriculture, health, and everyday conversation. The dataset includes eight transcriptions (two per language) from audio recordings ranging from 18 seconds to 2 minutes 34 seconds. This demo presentation outlines key insights to inform further prompt creation and the analysis of over 3,000 hours of audio data for a language technology project currently in progress. Findings reveal that shorter recordings (18–29 seconds) tend to have high TTR scores, as they contain fewer tokens and show limited lexical variety, whereas longer recordings (above 1 minute) provide more stable diversity measures. This observation aligns with Fergadiotis et al. (2015), who highlight the importance of text length in assessing lexical diversity indices.

To ensure transcription accuracy, ASR tools were tested. The IgboSpeech ASR platform was used to transcribe a sample Igbo audio of 22 seconds, and while the overall proficiency was good, the transcription contained spelling and tonal mark errors, which required manual corrections for accuracy. The Spitch software, on the other hand, failed to transcribe beyond 15 seconds of audio and produced errors in Yoruba transcriptions, confirming that manual transcription remains more reliable for under-resourced languages. Lexical diversity varied by discourse type. Instructional speech, such as maternal healthcare discussions, had higher MTLD scores due to structured vocabulary use, while conversational speech exhibited lower diversity with formulaic expressions.

- Agriculture (Igbo & Hausa) – Process-Oriented Speech
Agricultural discourse focused on food production. An 18-second Igbo transcript on corn processing had a TTR of 0.759 (29 tokens, 22 unique), with repeated terms like *oka* (‘corn’) and *ese* (‘grind’). A 29-second transcript on pepper harvesting had a TTR of 0.8421 (38 tokens, 32 unique), where words like *ose* (‘pepper’) and *gba* (‘pluck’) appeared frequently. The brevity of these recordings contributed to higher TTR scores. The 1:15-minute Hausa transcript on rice pounding had a TTR of 0.667 and an MTLD of 59.2558 (96 tokens, 64 unique). Repeated technical terms, such as *shinkafa* (‘rice’) and *buga* (‘pound’), reduced lexical diversity.

- Health (Yoruba & Hausa) – Instructional Speech
Health-related transcripts (1:40–2:34 minutes) featured structured advisory content. A Yoruba maternal health transcript (1:40 minutes) had a TTR of 0.484 and an MTLD of 49.4323 (219 tokens, 106 unique). Common advisory terms included *Iya* (‘mother’) and *ọmọ* (‘child’). Another Yoruba transcript on mental health (1:53 minutes) had a TTR of 0.5203 and an MTLD of 53.6948 (123 tokens, 64 unique), often repeating *awon* (‘they’) and *eniyan (person)*. Also, the Hausa immunisation transcript (2:34 minutes) had a TTR of 0.5246 and an MTLD of 53.9092 (183 tokens, 96 unique), frequently using *rigakafi* (‘vaccination’), *ciki* (‘pregnancy’), and *lafiya* (‘health’). While length increased lexical stability, the repetition of medical terms constrained variation.

- Everyday Conversation (Nigerian Pidgin) – Informal Speech
Nigerian Pidgin transcriptions relied on discourse markers (*wey, dey, na*), leading to the lowest diversity scores (TTR: 0.40–0.49, MTLD: 23.36–30.12). A 1:32-minute transcript had a TTR of 0.4969 and an MTLD of 26.0607 (159 tokens, 79 unique), frequently using *abeg* (‘please’). Another (1:36 minutes) had a TTR of 0.4221 and an MTLD of 23.3601 (154 tokens, 65 unique), dominated by *wetin* (‘what’).

Comparison of Lexical Diversity Metrics Across Languages

The following table summarises TTR, MTLD, total tokens, and unique tokens for each language across domains:

Language	Domain	Recording Length	Total Tokens	Unique Tokens	TTR	MTLD
Igbo	Agriculture - Corn Processing	0:18	29	22	0.759	27.36
Igbo	Agriculture - Pepper Harvesting	0:29	38	32	0.842	67.39
Yoruba	Health - Maternal health	1:40	219	106	0.484	49.43
Yoruba	Health - Mental health	1:53	123	64	0.52	53.69
Hausa	Agriculture - Rice pounding	1:15	96	64	0.667	59.26
Hausa	Health	2:34	183	96	0.525	53.91
Nigerian Pidgin	Everyday Conversation	1:32	159	79	0.497	26.06
Nigerian Pidgin	Everyday Conversation	1:36	154	65	0.422	23.36

These findings align with Sands (2009), who emphasises that African languages exhibit structural complexity due to their diverse typological features. It highlights the need for refined corpus-building approaches in low-resource languages. By integrating image, video, and text-based prompts, the study underscores the role of contextual variation in shaping lexical diversity, with implications for AI training data, linguistic theory, and multilingual speech recognition systems.

Keywords: Lexical Diversity; African Languages; Corpus Linguistics; Contrastive Linguistics; Computational Linguistics; AI in Linguistic Analysis

References

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). Journal of Quantitative Linguistics, 17(2), 94–100. doi: 10.1080/09296171003643098

Jarvis, S., & Hashimoto, B. J. (2021). How operationalizations of word types affect measures of lexical diversity. International Journal of Learner Corpus Research, 7(1), 163–194. doi: 10.1075/ijlcr.20004.jar

Fergadiotis, G., Wright, H. H., & West, T. M. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. Journal of Speech, Language, and Hearing Research, 58(3), 840–852. doi: 10.1044/2015_JSLHR-L-14-0280

Sands, B. (2009). Africa's Linguistic Diversity. Language and Linguistics Compass, 3(2), 559–580. doi: 10.1111/j.1749-818x.2008.00124.x

Globalization Partners. (2025). African Languages: Diversity, Classification, and Linguistic Features. Retrieved from : https://www.globalizationpartners.com/2025/02/21/african-languages-diversity-classification-challenges/

Wikipedia (2025). Languages of Nigeria. In Wikipedia, The Free Encyclopedia. Retrieved from https://en.wikipedia.org/wiki/Languages_of_Nigeria.