A Morpho-Syntactically Annotated Corpus of Ògè Folk Narratives with a Focus on Nominal Structure

Priscilla Lola Adenuga
Independent Researcher
Wiesbaden, Germany
priscillalola_adenuga@yahoo.com

Abstract

This paper presents a manually annotated morpho-syntactic corpus of Ògè, an under-resourced indigenous language spoken in Nigeria. The corpus consists of ten folk narratives (approximately 4,667 tokens) collected for the investigation of nominal structure. Annotation is expert-driven and includes token-level part-of-speech tagging together with a structured Determiner Phrase (DP) classification framework designed to capture language-specific nominal configurations. The scheme distinguishes between bare nouns and modified noun phrases, reflecting a central structural property of Ògè: noun forms remain morphologically stable across contexts, while modifiers exhibit formal and positional variation contributing to reference, specificity, and discourse prominence. The DP classification layer encodes both simple and complex nominal constructions, enabling systematic analysis of internal phrase structure. Designed as a reusable digital resource, the corpus supports morphosyntactic tagging, noun phrase boundary detection, and modeling of nominal structure in low-resource NLP settings. The annotated dataset will be made publicly available through the SADiLaR repository. This work demonstrates how descriptive linguistic analysis can inform annotation design and provides a replicable framework for developing structured resources for under-resourced African languages.

Keywords: Ògè, low-resource NLP, annotated corpus, nominal structure, African languages

1. Introduction

Many African indigenous languages remain severely under-resourced with respect to publicly available linguistic data and computational tools (Joshi et al., 2020; Martin et al., 2022). The absence of structured, annotated corpora limits both descriptive linguistic research and the development of Natural Language Processing (NLP) systems for these languages. The need for linguistically informed resource creation has been widely recognized as essential for addressing this gap (Joshi et al., 2020; Martin et al., 2022).

Ògè, an indigenous language spoken in Nigeria and belonging to the Benue-Congo family, exemplifies these challenges. While typologically related languages such as Yorùbá have received increasing attention in corpus-based and computational research, Ògè lacks digitally accessible annotated resources. Developing structured resources for Ògè is therefore important not only for documentation purposes, but also for expanding the typological and computational coverage of smaller African indigenous languages within the same linguistic space. Despite exhibiting a rich nominal system involving nouns, pronouns, determiners, numerals, and modifiers, Ògè lacks annotated corpora that can support systematic linguistic analysis or computational modeling. In narrative discourse, nominal expressions occur both as bare nouns and as modified noun phrases, raising important questions about the structural encoding of reference, specificity, and discourse prominence. These properties require annotation strategies that are sensitive to language-internal structure rather than directly inherited from high-resource language frameworks. This paper presents a morpho-syntactically annotated corpus of ten Ògè folk narratives designed to capture nominal structure in naturally occurring discourse. The corpus contains approximately 4,667 tokens and is manually annotated with token-level part-of-speech tags and a structured Determiner Phrase (DP) classification framework. A central feature of the annotation scheme is the explicit distinction between bare nouns and modified noun phrases, reflecting the morphological stability of noun forms and the structural variation of modifiers. By integrating descriptive linguistic analysis with structured annotation, this resource provides foundational infrastructure for both linguistic research and future NLP applications in low-resource settings. More broadly, the work contributes to ongoing efforts to develop high-quality, reusable, and linguistically grounded language resources for African indigenous languages.

2. Linguistics Background on Ògè

Ògè is an indigenous language spoken in Nigeria and belongs to the Benue-Congo language family. It shares typological properties, including SVO word order, with better-documented regional languages within the same linguistic space. Unlike these languages, however, Ògè lacks digitally accessible annotated resources, highlighting the need for structured resource development tailored to smaller indigenous languages.

The present study focuses specifically on the nominal system of Ògè, which plays a central role in the encoding of reference and participant tracking in narrative discourse. Nominal

Proceedings of Resources for African Indigenous Languages (RAIL) 2026 @ LREC 2026, pages 1–6
12 May 2026.

expressions in Ògè include nouns, pronouns, determiners, numerals, and adjectival modifiers. A salient structural property of the language is that nouns may occur either as bare forms or within modified noun phrases. Crucially, the morphological form of the noun remains stable across these contexts: nouns do not undergo overt inflectional change when appearing with or without modifiers. In contrast, nominal modifiers exhibit formal and positional variation depending on syntactic and semantic environment. Bare nouns are frequently attested in narrative discourse and are fully interpretable in context. Modified noun phrases introduce additional descriptive, quantitative, or contrastive information that refines interpretation. While modifiers are not obligatorily required for referential interpretation, their distribution reflects discourse-level considerations such as specificity, emphasis, and differentiation among participants. These structural characteristics make the nominal domain in Ògè particularly suitable for targeted annotation and systematic corpus-based investigation.

3. Related Work

The development of annotated language resources has been widely recognized as a prerequisite for both descriptive linguistic research and Natural Language Processing, particularly for African indigenous and other under-resourced languages (Joshi et al., 2020; Martin et al., 2022). Several efforts have focused on participatory and linguistically informed approaches to African language resource development (Nekoto et al., 2020). Previous work has emphasized that annotation frameworks designed for high-resource languages do not always adequately capture the structural properties of under-resourced languages (Bird, 2009). Large multilingual initiatives such as Universal Dependencies provide valuable cross-linguistic infrastructure, but language-specific adaptation remains essential (Zeman et al., 2020). Despite these efforts, available corpora for African indigenous languages remain limited in annotation depth, particularly with respect to fine-grained nominal distinctions. The present work contributes to this area by offering a nominally focused annotated corpus of Ògè narratives grounded in descriptive linguistic analysis.

4. Corpus Description

The corpus consists of ten Ògè folk narratives collected as part of long-term descriptive research on nominal structure. The texts belong to the genre of oral folklore and reflect culturally embedded narrative traditions within the speech community. They were transcribed from recorded speech and subsequently prepared for linguistic annotation. The present corpus represents a curated and systematically annotated subset of a broader body of Ògè data compiled over several decades of fieldwork. The selection of ten narratives reflects a balance between data quality and annotation feasibility, as the corpus was manually annotated using an expert-driven approach. Texts were selected based on transcription quality, completeness, and representativeness of narrative discourse.

The annotated dataset contains approximately 4,667 tokens across ten narratives, with an average of approximately 467 tokens per text. The narratives vary in length and include both descriptive passages and dialogic segments, thereby providing diverse discourse environments in which nominal expressions occur. Nominal tokens (nouns and pronouns) constitute a substantial proportion of the corpus, reflecting the narrative emphasis on participants and reference tracking. This distribution provides a rich empirical basis for examining the contrast between bare nouns and modified noun phrases in naturally occurring discourse.

Transcription followed a consistent orthographic representation aligned with established descriptive conventions for Ògè. Minor normalization was performed to ensure cross-text consistency in cases of orthographic variation. No structural simplification or artificial modification of the narratives was introduced. Although elicited examples were consulted during linguistic analysis to clarify contrasts in nominal modification, the annotated corpus itself contains exclusively naturally occurring narrative data. Elicited materials were deliberately excluded to preserve the integrity of spontaneous discourse within the dataset. Transcription was carried out by trained speakers of the language and reviewed for consistency. Annotation decisions were iteratively checked to ensure reliability across texts.

5. Annotation Scheme

The corpus is annotated manually using an expert-driven approach informed by prior descriptive analysis of Ògè nominal structure. Annotation was conducted at the token level and includes part-of-speech tagging across all word classes. Tokens were segmented according to orthographic word boundaries established during transcription. The annotated data are organized in a structured tabular format, with each token associated with its corresponding part-of-speech (POS) tag and DP classification label. The annotation scheme assigns part-of-speech categories to nouns, pronouns, determiners, adjectives, verbs, and other functional elements. While all tokens receive POS tags to preserve contextual completeness, the primary analytical focus of the annotation lies in nominal categories, namely nouns, pronouns, determiners, and nominal modifiers.

A central feature of the scheme is the explicit marking of bare nouns in contrast with modified noun phrases. Bare nouns are identified as nominal heads occurring without overt modification within their local syntactic context. Modified noun phrases are annotated with clear marking of modifier elements and their structural relation to the head noun.

Because noun forms in Ògè remain morphologically stable across contexts, special attention was devoted to identifying modifier variation and ensuring consistent categorization. Annotation decisions were guided by descriptive criteria established during prior linguistic analysis. Edge cases were reviewed iteratively across texts to maintain consistency. A written annotation guideline was developed to ensure systematic application of category definitions throughout the corpus. Elicited examples derived from the same narratives were consulted during annotation to resolve ambiguous cases and confirm structural interpretations. However, these elicited materials were not incorporated into the corpus itself. This approach ensured that annotation remained grounded in natural discourse while benefiting from controlled linguistic validation.

5.1 DP Classification Framework

In addition to part-of-speech tagging and nominal-level annotation, the corpus incorporates a structured classification framework for Determiner Phrases (DPs). This framework builds on established glossing and abbreviation conventions, particularly the Leipzig Glossing Rules and related proposals for morphological annotation (e.g., Corbett, 2000; Corbett, 2006; Creissels, 2006), while adapting them to the structural properties of Ògè nominal expressions. The classification system distinguishes between simple and complex DP configurations attested in the narratives. Categories include independent pronouns, bare nouns, demonstratives, quantifiers, numerals (cardinal and ordinal), and multi-element nominal constructions. Complex DPs are encoded through compositional labels that reflect their internal structure, such as:

- N (Noun)
- BARE N (Bare Noun)
- N + A (Noun + Adjective)
- N + PRO (Noun + Pronoun)
- D + N (Demonstrative + Noun)
- CARD + N (Cardinal + Noun)
- Q + N (Quantifier + Noun)
- N + N (Noun + Noun)
- Multi-element structures (e.g., D + N + N + PRO)

To further illustrate the annotation scheme, Table 1 provides representative examples of selected DP configurations observed in the corpus.

DP Type	Example	Gloss
N + A	ísinsìn ópú	‘black dog’
D + N	ígé ópú	‘that/those house’
CARD + N	íkín ópú	‘one dog’
N + PRO	uwan’rin	‘your child’

Table 1: Examples of DP configurations in the Ògè corpus

The use of systematic abbreviation conventions ensures consistency with broader morphological annotation practices while allowing sufficient granularity to capture language-specific DP patterns. Rather than imposing externally defined syntactic templates, the scheme reflects distributional regularities observed in the corpus. This DP classification layer complements token-level POS tagging by providing structural information about nominal phrase composition. It enables quantitative analysis of DP complexity and facilitates future computational modeling of nominal structure in Ògè.

5.2 Distribution of DP Types in the Corpus

The DP classification framework enables quantitative observation of nominal phrase configurations across the corpus. Preliminary inspection of the annotated data indicates that bare nouns and simple noun-modifier constructions are among the most frequent DP types. Multi-element structures involving demonstratives, numerals, pronouns, and adjectives also occur, particularly in descriptive or contrastive contexts within the narratives. While the corpus was not designed as a statistical survey of DP frequency, the annotation reveals systematic distributional patterns. Bare nouns frequently appear in contexts where referents are discourse-given or contextually recoverable. Modified DPs tend to occur in environments requiring disambiguation, emphasis, or descriptive elaboration. The presence of structurally complex DPs, including multi-element configurations (e.g., demonstrative + noun + pronoun), demonstrates that Ògè nominal structure exhibits internal layering that can be systematically captured through annotation. This distributional diversity reinforces the value of incorporating DP-level classification in addition to token-level POS tagging.

Future quantitative analysis of the corpus may further explore correlations between DP complexity and discourse function, thereby extending the resource's applicability to computational modeling of nominal phrase structure.

6. Nominal Modification and Bare Nouns in Ògè

Nominal expressions in Ògè narratives occur both as bare nouns and as modified noun phrases. A central structural property observable in the corpus is the morphological invariance of noun forms across these contexts. Referential refinement is achieved through the addition of modifier elements while the nominal head remains formally stable. Bare nouns are frequent throughout the narratives and are fully interpretable in context. They typically occur where referents are discourse-given or contextually recoverable, and in such cases the nominal head alone suffices to identify the intended referent. Preliminary inspection of the corpus suggests that bare noun constructions constitute a substantial proportion of nominal occurrences, particularly in participant-tracking contexts within narrative discourse. Modified noun phrases, by contrast, introduce descriptive, contrastive, or quantificational information that refines interpretation. Nominal modifiers contribute to specificity, emphasis, and differentiation between entities. Their distribution is not arbitrary; rather, modifiers tend to occur in contexts where speakers aim to highlight salient properties, disambiguate reference, or signal narrative prominence. The data therefore suggest that nominal modification in Ògè is structurally layered rather than morphologically encoded: semantic and discourse distinctions are realized through compositional structure rather than head alternation.

From an annotation perspective, distinguishing between bare nouns and modified noun phrases required structural rather than surface-level criteria. Because noun forms do not vary morphologically across contexts, the presence or absence of modification must be determined through syntactic configuration and contextual interpretation. This structural approach informed the explicit encoding of bare nouns and modifier elements within the annotation scheme.

6.1 Illustrative Annotated Example

To illustrate the annotation scheme and the distinction between bare nouns and modified noun phrases, consider the following excerpts from the corpus.

Example (1)
Túndé rá vè ùwà.
Tunde FUT go farm
'Tunde will go to the farm.'

In this sentence, *ùwà* 'farm' functions as a bare noun. The DP consists solely of a nominal head without determiner, numeral, or adjectival modification. The noun retains its canonical morphological form, and interpretation relies on contextual familiarity within the narrative.

Example (2)
Í gbádun òtúro ìyí ìgú nì dọwẹ.
I enjoy two DET year FOC stay
'I enjoyed the two years that I stayed.'

Here, the nominal expression includes a quantificational modifier. The noun head appears in its stable morphological form, while the modifying element contributes additional semantic specification. Within the annotation scheme, the noun is tagged as NOUN, and the modifier is encoded as part of a structured DP configuration reflecting its internal composition.

Example (3)
Sadé pu ígbegbe íwaji ísinsin òdí.
Sade kill small bad black rat
'Sade killed a small bad black rat.'

In this example, the noun phrase contains multiple elements, illustrating a more complex DP configuration. The annotation captures both the nominal head and the internal structure of the modifying elements.

In the corpus, such contrasts are systematically marked: bare nouns are annotated as nominal heads without associated modifier tags, whereas modified noun phrases include explicit encoding of modifier elements and their structural relation to the head noun. These examples demonstrate how the annotation captures structurally meaningful variation while preserving head-form stability.

6.2 Annotation Challenges and Linguistic Insights

The annotation process revealed challenges that underscore the necessity of linguistically informed corpus design for under-resourced languages. As observed in prior work, annotation schemes developed for high-resource languages do not always transfer straightforwardly to structurally distinct systems (Bird, 2009; Zeman et al., 2020). In Ògè, the absence of overt morphological alternation between bare and modified contexts means that modification cannot be identified through surface morphology alone. Instead, structural configuration and discourse context must guide annotation decisions. One recurrent challenge involved distinguishing tightly integrated modifier constructions from potentially lexicalized sequences. In certain cases, modifier elements appear in close proximity to the noun and show reduced phonological or orthographic distinction, raising segmentation and categorical questions. Similar segmentation issues have been documented in annotation efforts for

morphologically rich and under-resourced languages (Bender, 2011). These cases required iterative cross-text comparison and consultation of elicited data to ensure consistency.

A further challenge concerned discourse-sensitive interpretation. Bare nouns may function differently depending on whether a referent is newly introduced, previously mentioned, or contextually salient. Although discourse status is not directly encoded in the annotation layer, annotation decisions were informed by narrative context to ensure accurate identification of nominal heads and modifier structures. The interaction between nominal structure and discourse prominence is central to reference modeling (Joshi et al., 2020), reinforcing the importance of structural precision in corpus annotation.

Taken together, these observations demonstrate that accurate annotation of Ògè nominal structure requires structural and discourse-sensitive criteria rather than purely morphological diagnostics. The resulting corpus reflects language-internal distinctions while remaining suitable for computational modeling and reuse in low-resource NLP contexts.

7. NLP Use Cases and Implications

The annotated corpus supports morphosyntactic tagging and nominal phrase identification in low-resource settings. Because DP boundaries and internal structure are explicitly encoded, the dataset can serve as supervised training material for noun phrase boundary detection, sequence labeling, and structural parsing experiments. The explicit marking of nominal heads and modifier elements provides structured input suitable for modeling internal DP composition and reference tracking.

As a linguistically grounded resource, the corpus provides a foundation for NLP applications involving Ògè and related African indigenous languages. The morphological stability of noun forms, combined with variation in modifier realization, provides a controlled environment for investigating nominal structure modeling in low-resource contexts. The explicit distinction between bare nouns and modified noun phrases may facilitate supervised or transfer-learning approaches to modeling nominal structure. Furthermore, the dataset may support comparative research across African indigenous languages exhibiting similar nominal patterns

8. Ethics, Limitations, and Data Availability

The corpus consists of culturally embedded folk narratives recorded within the speech community. The texts do not contain sensitive personal information, and no identifiable private data are included. The narratives represent traditional oral material rather than contemporary personal accounts. The resource is intended to support language documentation and digital preservation efforts while respecting community ownership of cultural knowledge.

The present release contains approximately 4,667 tokens and focuses specifically on nominal structure and DP-level annotation. While part-of-speech tagging is provided for all tokens, deeper syntactic dependency annotation and discourse-level features are not included at this stage. The annotated narratives represent an initial release drawn from a broader body of Ògè data collected over several decades, including additional narrative and elicited materials. Future work will extend the annotation framework to further texts within this larger archive. The annotated corpus will be deposited in the SADiLaR repository in a structured digital format suitable for reuse. The release will include the annotated texts together with documentation of the annotation scheme and DP classification framework to facilitate reproducibility and further research.

9. Conclusion

This paper has presented a manually annotated morpho-syntactic corpus of ten Ògè folk narratives (approximately 4,667 tokens), with a structured focus on nominal expressions. The resource combines token-level part-of-speech tagging with a linguistically grounded DP classification framework that explicitly distinguishes between bare nouns and modified noun phrases. By capturing both simple and complex nominal configurations, the corpus documents structural properties central to reference and discourse organization in Ògè.

The annotation design demonstrates how detailed descriptive linguistic analysis can inform the development of reusable digital resources for under-resourced languages. As a publicly available dataset, the corpus provides foundational infrastructure for computational work on Ògè, including morphosyntactic tagging, noun phrase boundary detection, and modeling of nominal structure in low-resource settings. More broadly, this work supports ongoing efforts to expand structured digital resources for African indigenous languages and to strengthen their presence in computational research.

10. Bibliographical References

Bender, E. M. (2011). On Achieving and Evaluating Language-independence in NLP. *Linguistic Issues in Language Technology*, *6*.

Bird, S. (2009). Last Words: Natural Language Processing and Linguistic Fieldwork. *Computational Linguistics*, *35*(3), 469-474.

Corbett, G. G. (2000). *Number*. Cambridge University Press.

Corbett, G. G. (2006). Agreement. Cambridge: Cambridge University Press.

Creissels, D. (2006). *Syntaxe générale une introduction typologique 1: catégories et constructions*.

Lehmann, C. (1982). Directions for Interlinear Morphemic Translations. *Folia Linguistica*, *16*.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020, July). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282-6293).

Orife, I., Kreutzer, J., Sibanda, B., Whitenack, D., Siminyu, K., Martinus, L., Ali, J. T., Abbott, J., Marivate, V., Kabongo, S., Meressa, M., Murhabazi, E., Ahia, O., Van Biljon, E., Ramkilowan, A., Akinfaderin, A., Öktem, A., Akin, W., Kioko, G., . . . Bashir, A. (2020). Masakhane – Machine Translation for Africa. *ArXiv*. https://arxiv.org/abs/2003.11529

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., ... & Bashir, A. (2020). Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In *Findings of the Association for Computational Linguistics, EMNLP* (pp. 2144-2160).

Nivre, J., De Marneffe, M. C., Ginter, F., Hajic, J., Manning, C. D., Pyysalo, S., ... & Zeman, D. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 4034-4043).