Nigerian Journal of Operations Research (NIJOR)

https://nijor.org.ng/
pISSN: 2141-7938
eISSN: 3156-5956

Development and Evaluation of an Annotated Nya-Həba Corpus for Named Entity Recognition

Chahyaandida Ishaya[1]*, Ayuba Madu Yami[2], Ikeme John Dike[3]

[1,2,3] Department of Operations Research, Modibbo Adama University, Yola, Adamawa State, Nigeria

ARTICLE INFO

Article history:
Received 06 October 2026
Revised form 11 November 2026
Accepted 15 November 2026
Available online 19 November 2026

Keywords: *Nya-Həba; Kilba; named entity recognition; corpus annotation; inter-annotator agreement; CRF; BiLSTM-CRF*

ABSTRACT

This paper reports the construction and evaluation of an annotated Nya-Həba corpus for named entity recognition (NER). Nya-Həba, also known as Huba or Kilba, is spoken by the Həba community in Hong Local Government Area, Adamawa State, north-eastern Nigeria. Texts from documentary and community sources were prepared, segmented and labelled in Label Studio. The resulting resource contains 2,038 annotation units, 27,123 tokens and 1,928 adjudicated entity mentions. The annotation covers people, places, temporal expressions and indigenous animal terms. Two fluent speakers labelled the same material independently using contextual beginning–inside–outside (BIO) rules; a third fluent speaker reviewed and resolved disagreements. Before adjudication, token agreement stood at 98.03%, with Cohen's $\kappa = 0.8631$ and exact entity-span/type F1 of 86.17%. The review identified 535 disputed tokens across 304 units. Three further personal-name labels on which the annotators had agreed were subsequently changed to O, and these corrections are recorded separately. Both original annotation files and the correction audit have been retained. Using a common split of approximately 70% training, 15% validation and 15% test data, the feature-based conditional random field (CRF) achieved 75.35% entity F1, compared with 68.19% for a bidirectional long short-term memory–conditional random field (BiLSTM-CRF) using randomly initialized embedding. The study's contribution is the documented corpus-building process, including source preparation, independent annotation, agreement analysis, adjudication and reproducible baseline experiments. Most of the material comes from biblical narrative because substantial, consistently written Nya-Həba materials are scarce, and these narrative books offer continuous contextual examples of persons, places and temporally situated events, while temporal and animal examples are relatively few. Written authorization was obtained from the relevant language development and Bible translation organizations. The corpus and reproducibility materials are available through Zenodo subject to the applicable source-use conditions.

* *Corresponding author.*
E-mail address: chahyaandidaishaya@gmail.com https://doi.org/10.31181/nijor/0304.031

1. Introduction

Named entity recognition (NER) is the task of locating expressions in a text and assigning them to categories such as people and places. It supports a range of information-extraction applications. Developing NER tools for languages with limited digital resources is difficult because relatively little annotated data, few language-processing tools and limited benchmark coverage are available. Work on African-language NER has begun to address this gap through community-led datasets and cross-lingual methods. MasakhaNER (Adelani *et al.*,

2021) and its expanded version, MasakhaNER 2.0 (Adelani *et al.*, 2022), are notable examples. Nya-Həba, however, has received little attention in computational language research. A documented corpus is therefore needed to provide a basis for examining annotation decisions and testing initial recognition models.

The present study responds to this need by developing an annotated Nya-Həba corpus from written materials and contributions made by members of the community in Hong Local Government Area (LGA), Adamawa State, Nigeria. The language is also referred to as Huba or Kilba and is listed under ISO 639-3 identifier hbb and Glottocode huba1236 (Glottolog, n.d.).

Four entity categories are used in the annotation: person names (PER), place names (LOC), time expressions (TIME) and animal terms in the indigenous language (ANIM). ANIM is retained as a separate category because the study covers ordinary animal terms as well as names referring to particular animals. This distinction is important when interpreting the scope of the resulting annotations.

Although African-language NER resources have increased in recent years, a comparable documented and independently annotated resource for Nya-Həba has not been available. In the absence of such a resource, it is difficult to assess annotation consistency or establish a baseline against which later systems can be compared. This work addresses that gap by documenting where the texts came from, how the labels were assigned, how disagreements were handled and how the baseline models were evaluated. The materials are made available for scrutiny and further work, subject to the conditions governing the original sources.

The work is guided by three questions. First, how were the Nya-Həba texts collected, prepared and annotated? Second, what level of agreement was reached by the two independent annotators before adjudication, and how were their differences resolved? Third, how well do a feature-based conditional random field (CRF) and a bidirectional long short-term memory network with a conditional random field output layer (BiLSTM-CRF) perform when trained and tested using the same data partition and evaluation procedure?

The main output is a four-category Nya-Həba NER corpus. Alongside the final annotated data, the study retains the two independent annotation files, reports agreement at token and entity levels, and provides an audit trail of corrections made during adjudication. The preparation procedure and annotation guideline are also documented. Finally, CRF and BiLSTM-CRF experiments are reported using a common data split, with the associated predictions, tables, figures and model files. The deposit identifier and materials are reported in the Data Availability Statement.

2. Literature Review

This section reviews prior work on the Nya-Həba linguistic setting, African-language named entity recognition, corpus annotation practices, and sequence-labelling baselines. It identifies the evidential gap addressed by a speaker-annotated, adjudicated corpus with transparent evaluation procedures.

2.1 Community and Dialect Scope

The study was conducted in Hong Local Government Area (LGA), Adamawa State, Nigeria, with the participation of fluent native speakers of Nya-Həba (Kilba). Muazu (2009) documents phonological and lexical correspondences between the Hong and Gaya varieties of the language, providing linguistic context for the speech community represented in this study.

The primary objective of this research was to construct and evaluate a Nya-Həba Named Entity Recognition (NER) corpus using a unified semantic annotation framework. Accordingly, the study focused on identifying and classifying named entities according to their contextual meanings rather than examining dialectal variation. Native-speaking contributors confirmed that the person (PER), location (LOC), animal (ANIM), and time (TIME) expressions examined generally maintain equivalent referential meanings across Hong and Gaya. Consequently, a consistent four-category annotation scheme was adopted throughout the corpus to ensure semantic uniformity and methodological consistency.

Four fluent native-speaking contributors from Hong LGA, familiar with both Hong and Gaya, participated in the linguistic consultation and corpus development process. Their linguistic knowledge supported the contextual interpretation and validation of named entities. The resulting corpus was therefore treated as a unified Nya-Həba language resource, consistent with the study's predefined objectives of corpus construction, annotation quality assurance, and computational evaluation.

Church and community translation initiatives are part of the history of written Nya-Həba. Balama's online account (2024) describes the involvement of Gabriel Fadin Amos, representing SUM(D), and Pastor Ishaku Kwambla of EYN in translation and orthographic activities said to have begun in 1971. The account also credits Bishop Amos Bukata Elisha with leading a team involved in translating the New Testament and hymn book. Since this is a community-oriented online source, it is used here only to provide historical context. The biblical material used in the corpus comes from the 2018 Nya-Həba New Testament, attributed to Wycliffe Bible Translators, Inc. and Nigeria Bible Translation Trust (Wycliffe Bible Translators & Nigeria Bible Translation Trust, 2018).

2.2 Related NER and Annotation Research

CoNLL-2003 helped establish the span-based approach and evaluation conventions widely used in named entity recognition (Tjong Kim Sang & De Meulder, 2003). For languages where manual annotation is scarce, WikiAnn demonstrated how multilingual name-tagging data could be assembled at scale through transferred or silver annotations, covering 282 languages (Pan *et al.,* 2017). These resources have widened the reach of NER research. At the same time, they show why carefully checked, language-specific annotations remain useful, particularly where local meanings, domain coverage and consistency of labelling need close attention.

Research on NER in Nigerian languages has also moved beyond general multilingual benchmarks. Alabi et al. (2020), for example, developed a manually annotated Yorùbá evaluation corpus while examining language representations for low-resource settings. MasakhaNER provided manually annotated news data for ten African languages, among them Hausa, Igbo, Nigerian Pidgin and Yorùbá (Adelani *et al.*, 2021). Its second release extended coverage to 20 languages and investigated Africa-centric transfer learning (Adelani *et al.*, 2022). Other work, including AfriBERTa (Ogueji *et al.,* 2021) and AfroXLMR (Alabi *et al.,* 2022), has explored language models and adaptation strategies suited to African languages. More recently, YoNER introduced a multi-domain Yorùbá dataset drawn from sources such as the Bible, blogs, films, radio broadcasts and Wikipedia (Falola *et al.*, 2026). Taken together, these studies provide useful precedents for the development of a Nya-Həba resource and point to the value of testing it beyond a single text domain. Experiments involving Hausa and isiXhosa have likewise shown that transfer learning and distant supervision can help multilingual transformers in low-resource settings, although performance depends on the language and experimental conditions (Hedderich *et al.*, 2020). Work on Runyankore further suggests that the choice of an auxiliary language can influence cross-lingual transfer (Asiimwe *et al.*, 2026).

Building a reliable corpus also requires more than assigning labels to words. Annotation projects need clear definitions, written instructions and a procedure for dealing with difficult cases. Independent annotation followed by adjudication is one established way of identifying and resolving differences (Pustejovsky & Stubbs, 2012). Agreement coefficients are useful, but they should be read in the light of the annotation task: they indicate how consistently annotators applied a coding scheme, not whether every interpretation is necessarily the correct one (Artstein & Poesio, 2008). For this reason, the present study retains both original annotation files, reports token- and entity-level agreement, and makes the project guideline available with the corpus.

The experiments use two established sequence-labelling approaches. The first is a linear-chain conditional random field (CRF), following Lafferty *et al.* (2001); the second is a bidirectional long short-term memory network with a CRF output layer (BiLSTM-CRF), following Huang *et al.* (2015). They are included as initial reference systems for the new corpus. The intention is not to claim that these models represent the strongest available methods, but to establish results that can be compared with later experiments using multilingual and African-focused pretrained models.

3. Methodology

The research followed a corpus-development and experimental-evaluation design. The stages comprised source selection, text preparation, native-speaker annotation, agreement assessment, adjudication, controlled partitioning and comparison of two sequence-labelling baselines. Ethical permission and reproducibility measures are also documented.

3.1 Sources and Collection

The texts were assembled between 21 June and 30 August 2026. As shown in Table 1, the prepared material was organised into four source partitions. Acts and Luke were selected deliberately because they contain many references to people and places and formed a substantial written source of Nya-Həba material available to the researcher. These narrative books offered continuous, consistently written contexts for person, location and temporal references in a language with limited accessible digital text. Biblical parallel corpora have previously supported low-resource language technology research (Christodoulopoulos & Steedman, 2015; Agić *et al.*, 2015). This precedent supports resource construction, not any inference of contemporary-domain NER performance. The remaining 390 units (19.14%) came from the other three partitions.

Table 1: Sources of the prepared annotation units

Source	Contributor or document	Units	Share (%)
Acts and Luke	Nya-Həba New Testament (2018) (Wycliffe Bible Translators and Nigeria Bible Translation Trust 2018)	1,648	80.86
Personal-name material	Richard Dauda (Maksha Gong)	98	4.81
Oral narrative	Rev. Simon Yohanna	177	8.68
Folktale-derived material	Mrs. Asenah Abama	115	5.64
Total	Four prepared-text partitions	2,038	100.00

Source: *Authors' prepared Nya-Həba corpus and source-provenance records (2026); Acts and Luke derive from Wycliffe Bible Translators and Nigeria Bible Translation Trust (2018).*

The personal-name material was provided by Richard Dauda (Maksha Gong), a traditional language teacher and assistant pastor from Tapchima in Hong LGA. Rev. Simon Yohanna contributed the oral narrative, while Mrs. Asenah Abama narrated the folktales used to prepare the additional material. The four partitions counted in Table 1 are the materials prepared for annotation.

3.2 Preparation and Representation

The Nya-Həba textual materials were extracted, standardized, segmented, and consolidated for manual annotation. The corpus combined biblical passages from Acts and Luke with personal names, oral narratives, and additional language materials. The accompanying archive documents source verification in data_cleaning/produce_1648_acts_luke_sen.py and text aggregation in data cleaning/build_sentences_full.py.

The preprocessing pipeline involved Unicode NFC normalization, standardization of curly apostrophes and backticks to straight ASCII apostrophes, casefolding, removal of numerical references, punctuation handling, and whitespace normalization. During source verification, apostrophes were preserved within words, while hyphens were treated as word separators. The prepared annotation text contained no leading hyphenated words. The original surface-text segments were retained for annotation rather than replaced with their normalized matching representations.

The consolidated dataset comprised 2,038 annotation units, each limited to 18 whitespace-delimited words, with 705 units reaching this limit. Each unit was structured as a JSON object containing a text field and imported into Label Studio version 1.23.0. Named entities were manually assigned four labels—person (PER), location (LOC), animal (ANIM), and time (TIME)—based on contextual meaning. Punctuation marks were treated as non-entity elements and assigned the outside (O) tag in the BIO representation.

Following annotation, the dataset was represented in CoNLL-2003 format using the Beginning–Inside–Outside (BIO) tagging scheme, producing 27,123 tokens across 2,038 annotation units, with a maximum of 25 tokens per unit. The independent annotator outputs and adjudicated corpus were retained separately to support inter-annotator agreement analysis and annotation quality assurance. Validation confirmed consistent

annotation units and token sequences, with no malformed records or invalid BIO sequences. The final adjudicated corpus is distributed as conll/nya_heba_strict_full.conll.

3.3 Annotators and Procedure

Both annotators were from Hong LGA and were fluent speakers of Nya-Həba. Before the exercise, the study supervisor introduced them to the project rules and provided training using worked examples. They then labelled the same material independently and did not have access to each other's decisions while annotating. A third fluent speaker of Nya-Həba reviewed the differences and acted as adjudicator. All agreement figures reported in this section are based on the two original annotation files, before adjudication.

The project guideline, titled Nya-Həba NER Annotation Guideline, sets out how the labels were to be applied. Annotators worked with the supplied tokens as they appeared; they were not permitted to split, combine or retype them. The scheme uses nine BIO labels: B-PER, I-PER, B-LOC, I-LOC, B-TIME, I-TIME, B-ANIM, I-ANIM and O. Context was central to the decision. A word could be tagged as an entity in one annotation unit and labelled O in another if its use differed. For a name made up of several tokens, the first token received a B- label and the remaining tokens received I- labels. A new entity in a list or beside another entity began with a fresh B- label. Repeated words were considered separately, punctuation was labelled O, and uncertain cases were referred to the adjudicator rather than assigned a label by guesswork. Table 2 presents the four entity categories. ANIM covers animal terms in the indigenous language, while TIME includes calendar-related expressions and festivals when they function as time references in context. The use of explicit definitions and an adjudication procedure follows established practice in corpus construction (Pustejovsky & Stubbs, 2012).

Table 2: Annotation categories and final entity counts

Category	Operational definition	Final mentions
PER	Names of people.	1,004
LOC	Names of places, villages and cities in context.	627
TIME	Native-language time expressions: months, days, weeks, years and festivals.	202
ANIM	Animal names or animal terms in the native language.	95
O	Tokens outside the four entity categories.	Not an entity
Total	Entity mentions across the four categories.	1,928

Source: *Authors' adjudicated Nya-Həba NER corpus* (Zenodo, 2026)

3.4 Agreement Measures

Agreement was examined before adjudication using three complementary measures. Token agreement is the proportion of aligned tokens for which both annotators assigned exactly the same BIO label, including O. Cohen's kappa (κ) adjusts the observed agreement for the level of agreement that might be expected from the annotators' label distributions (Cohen, 1960). Its interpretation follows guidance on agreement measurement in computational linguistics (Artstein & Poesio, 2008). The coefficient is calculated as $\kappa = (p_o - p_e)/(1 - p_e)$, where p_o represents observed agreement and p_e is the sum of the products of the two annotators' marginal proportions across the labels. BIO prefixes were retained as part of each label in the primary kappa calculation.

Entity-level agreement was assessed more strictly. An entity counted as an exact match only when both annotators identified it in the same unit, with identical start and end boundaries and the same entity type. If A and B represent the sets of entities identified by the two annotators, the symmetric exact-match F1 is calculated as $F1 = 2|A \cap B|/(|A| + |B|)$. This calculation gives equal weight to both annotators and excludes O labels. It therefore captures differences in entity inclusion and span boundaries that may not be apparent from overall token agreement alone.

The two annotators assigned identical BIO labels to 26,588 of the 27,123 tokens. This corresponds to 98.03% token agreement and a Cohen's κ of 0.8631. The remaining 535 tokens, located in 304 annotation units,

received different labels; 1,734 units had complete agreement. When the BIO prefixes were removed and agreement was calculated using only the entity categories, κ was 0.8645. At entity level, 1,685 spans matched in both boundary and type. Chahyaandida identified 1,801 entities, while Ndachimya identified 2,110, giving an overall exact-match F1 of 86.17%. The breakdown by entity category is presented in Table 3.

Table 3: Pre-adjudication exact entity-span/type agreement

Type	Chahyaandida	Ndachimya	Exact matches	F1 (%)
PER	1,004	1,188	898	81.93
LOC	572	625	568	94.90
TIME	130	202	124	74.70
ANIM	95	95	95	100.00
All	1,801	2,110	1,685	86.17

Source: *Authors' independent annotation files and pre-adjudication agreement calculations (2026).*

The lowest exact entity agreement occurred for TIME, followed by PER. The difference between the annotators' total entity counts also indicates that they did not always select the same expressions for tagging. By contrast, all 95 ANIM mentions identified in the two files matched exactly, producing 100% agreement for that category in this dataset. This result should be understood in relation to the examples available: it does not establish that animal terms will be equally straightforward in other texts or contexts. In the present corpus, the main areas of annotation disagreement were temporal expressions and personal names.

3.5 Correction Audit

The adjudication stage focused on the 535 token positions where the two original annotations differed. Final labels were assigned in accordance with the contextual rules in the project guideline. The adjudication CSV records the sentence_id, full_sentence, token, your_tag, Ndachimya_tag, final_tag and adjudication_note fields. In this file, the field named your_tag refers to Chahyaandida's annotation. A separate token-position file was retained to distinguish repeated occurrences of the same word within an annotation unit. The full_sentence field is reconstructed from the aligned token sequence and represents the complete annotation unit.

Of the 535 disputed positions, the final decision followed Chahyaandida's label in 391 cases and Ndachimya's label in 144 cases. No third label was introduced for these disputed positions. The adjudicator also reviewed three occurrences of gəna that both annotators had labelled B-PER and changed them to O. These occur in units 579, 667 and 1494. Because the two annotators had initially agreed on these labels, the three changes are recorded separately from the 535 disagreements. The correction record therefore distinguishes disagreement resolution from corrections made after the initial agreement.

3.6 Experimental Split and Scoring

The two models were trained and tested using the same adjudicated corpus and the same data partition. Before splitting, annotation units were grouped by their NFC-normalised and case-folded token sequences so that exact duplicates, after normalisation, would remain in one partition. All 2,038 units in the current corpus were unique under this check. A fixed random seed of 42 was used to allocate approximately 70% of the units to training, 15% to validation and 15% to testing, as summarised in Table 4.

Table 4: Experimental split and entity supports

Partition	Units	Tokens	PER	LOC	TIME	ANIM
Training	1,426	18,739	703	452	134	70
Validation	305	4,116	136	95	33	13
Test	307	4,268	165	80	35	12
Total	2,038	27,123	1,004	627	202	95

Source: *Authors' adjudicated Nya-Həba corpus and reproducible split manifest (seed 42; 2026).*

The validation set was used for model selection, with entity F1 as the selection measure. On the test set, precision, recall and F1 were calculated using exact matches of both entity span and type. The scores are micro-averaged across entity categories, excluding O, and follow strict IOB2/CoNLL-style evaluation as implemented in seqeval (Tjong Kim Sang & De Meulder, 2003). Token accuracy, in contrast, counts exact BIO-label matches and includes O. Model predictions were evaluated as generated; no automatic BIO correction was applied after prediction. Neither model produced an invalid BIO transition on the test set. The experiments used seed 42 and were run on a CPU.

3.7 Feature-Based CRF

The feature-based linear-chain CRF was implemented with sklearn-crfsuite, using the structured prediction approach described by Lafferty et al. (2001). Parameter optimisation was performed with L-BFGS for a maximum of 100 iterations. The feature set included the normalised token, indicators of letter case, prefixes and suffixes of one to three characters, token length, digit and apostrophe indicators, information from neighbouring tokens, and markers for the beginning and end of each unit. No part-of-speech features were included, and all possible label transitions were allowed. Three equal L1 and L2 regularisation settings were tested ($c_1 = c_2 =$ 0.01, 0.1 and 1.0). The setting $c_1 = c_2 = 0.1$ gave the best validation entity F1, at 74.76%, and was selected for the final test.

3.8 BiLSTM-CRF

The second baseline was a neural sequence-labeling model consisting of one bidirectional LSTM layer and a linear-chain CRF decoder, following Huang et al. (2015). It used 100-dimensional word embeddings with random initialisation. The vocabulary was constructed only from the normalised, lowercased training tokens; separate indices were reserved for padding and unknown words. Each direction of the LSTM contained 128 hidden units, and dropout was set to 0.5. Packed sequences and masks were used so that padding tokens did not contribute as observed input. The implementation did not include a character encoder or pretrained word or contextual embeddings.

The model was trained with Adam at a learning rate of 0.001, using batches of 32 and gradient clipping at 5. Training was allowed to run for up to 30 epochs, with early stopping after five consecutive epochs without an improvement in validation entity F1. Training ended at epoch 24, and the checkpoint from epoch 19 was restored because it had the best validation score. That score was 65.65%. These details describe the implementation evaluated in this study; the full name BiLSTM-CRF is used throughout to avoid confusion with the less specific abbreviation 'Bi-CRF'.

3.9 Ethics, Access and Reproducibility

Written permission to conduct the research and use Həba-language documents, including biblical texts, was provided in an authority letter dated 26 October 2026 and signed on 27 October 2026. The letter was signed by Evang. Yahi Emmanuel Alaji Gaya, Secretary of the Həba Language Development Association, and Adulea B. Bulama, Coordinator of Nya-Həba Bible Translation, on behalf of Nigeria Bible Translation Trust, Jos. It was issued in response to the researcher's request of 13 October 2026 and records authorization from the Bible translation body, the language development association and the Həba community. A copy is held with the

project records and can be made available to the editor. The 2018 New Testament identifies Wycliffe Bible Translators, Inc. and Nigeria Bible Translation Trust as rights holders (Wycliffe Bible Translators & Nigeria Bible Translation Trust, 2018).

The dataset and accompanying project materials are deposited in a public repository. The repository is identified Zenodo, (2026), subject to the source-use conditions stated below.

The project archive includes the two original annotator files, the corrected CoNLL corpus, code used to calculate agreement, the annotation guideline, preprocessing scripts and the baseline implementations. Running the baseline script produces the data-split manifests, result tables, figures, token-level predictions, model files and a run manifest. The manifest records the random seeds, model settings, hashes of the input data and scripts, and the software package versions. The reported run used Python 3.12.14, sklearn-crfsuite 0.5.0, PyTorch 2.14.1+cpu, pytorch-crf 0.7.2 and seqeval 1.2.2. These retained materials make it possible for other researchers to inspect the data preparation and repeat or extend the experiments.

4. Results and Discussion

This section presents the held-out benchmark results for the conditional random field and bidirectional long short-term memory–conditional random field baselines, followed by an interpretation of the outcomes, class-level differences, corpus limitations and implications for further research.

4.1 Baseline Test Results

Table 5 compares token accuracy and exact entity-level precision, recall and F1 on the shared held-out test set. Each result is calculated from model predictions against the adjudicated reference annotations.

Table 5: Held-out test results. All values are percentages

Model	Token accuracy	Entity precision	Entity recall	Entity F1
CRF	96.77	77.54	73.29	75.35
BiLSTM-CRF	96.11	70.04	66.44	68.19

Source: *Authors' CRF and BiLSTM-CRF held-out test experiments (2026).*

On the held-out test set, the CRF obtained an entity F1 of 75.35%, which was 7.16 percentage points higher than the 68.19% recorded by the BiLSTM-CRF. For both models, precision was higher than recall, indicating that the systems were more successful at limiting incorrect entity predictions than at finding every entity present. Token accuracy was above 96% in both cases. That figure should be interpreted cautiously, however, because many tokens do not belong to an entity; high token accuracy does not necessarily mean that entity spans have been identified completely.

Table 6: Entity F1 by category on the test set

Entity type	Test mentions	CRF F1 (%)	BiLSTM-CRF F1 (%)
PER	165	83.54	77.96
LOC	80	77.02	67.46
TIME	35	38.71	36.36
ANIM	12	35.29	28.57

Source: *Authors' exact span/type evaluation on the held-out test set (2026).*

For both models, PER produced the highest category-level F1, with LOC in second place. Results for TIME and ANIM were lower. The test set contained 35 TIME mentions and only 12 ANIM mentions, so the scores for these categories are based on relatively few examples. Each model correctly identified three of the 12 ANIM mentions, equivalent to 25% recall. The contrast is noteworthy: annotators agreed on all 95 ANIM mentions in

the full corpus, but the models found only a quarter of the ANIM mentions in the test subset. Figure 1 presents the category-level results.

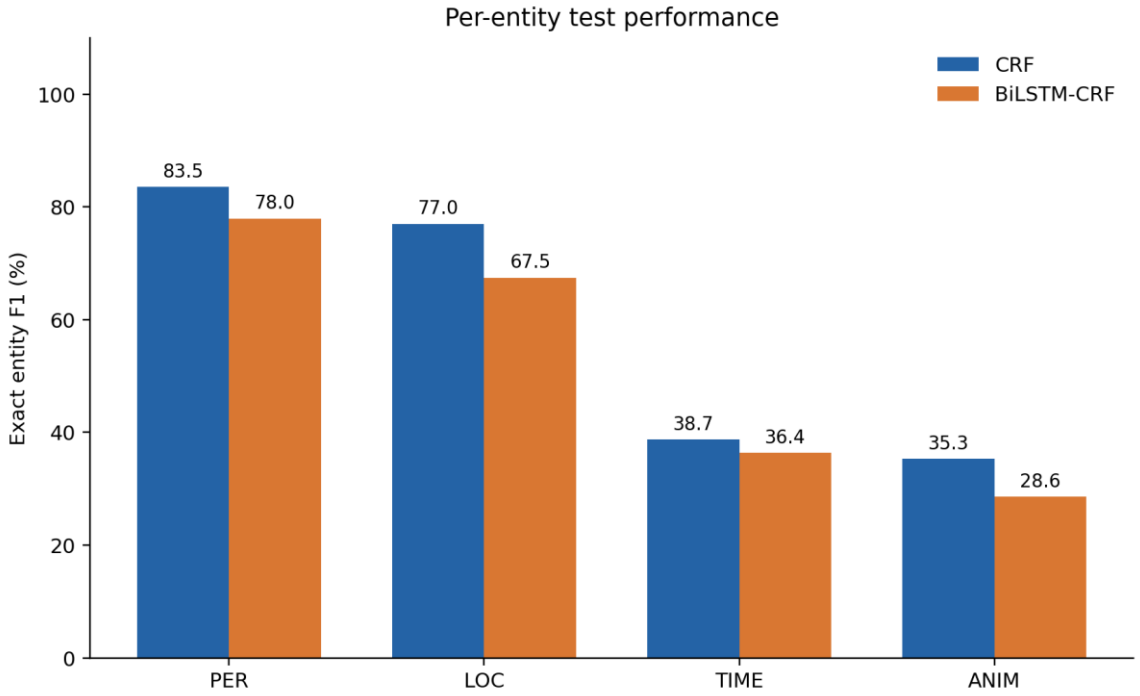

Fig. 1. Exact entity-span/type F1 by category for the shared held-out test set
Source: *Authors' test-set model predictions and entity-level evaluation (2026); values correspond to Table 6.*

4.2 Discussion and Limitations

The agreement figures and the model scores describe two different aspects of the corpus. The pre-adjudication kappa and exact entity F1 indicate how consistently the two annotators applied the agreed labelling rules. Test-set F1, on the other hand, shows how well a model trained on the corrected data identifies entities in units it has not seen during training. High agreement is useful evidence that the annotation procedure was applied consistently, but it should not be taken as proof that every label captures an unquestionable semantic interpretation (Artstein & Poesio, 2008). The ANIM results make this distinction clear: the annotators matched on all observed animal mentions, yet model recall for this category was 25% on the small test subset.

The CRF performed 7.16 percentage points better than the BiLSTM-CRF on the test split. This finding is limited to the particular split and random seed used in the experiment. One possible reason is that the CRF can make direct use of local word and spelling features, whereas the neural model had to learn its word representations from a relatively small training set, starting from random embeddings. The present experiment does not establish that this is the cause of the difference. Pretrained multilingual encoders, character-level models and African-focused transformers were outside the scope of the comparison. Further experiments using several random seeds would be needed before making a broader claim about the relative merits of the two model families.

The concentration of the source material is a major limitation. Acts and Luke account for 80.86% of the prepared units, meaning that biblical vocabulary and narrative style have a strong influence on the corpus. The community-contributed narratives and name material add useful variety, but they do not provide a balanced representation of different language-use settings. Since the models were not tested on contemporary news, ordinary conversation or other genres, their performance in those settings remains unknown. A future evaluation that withholds complete sources during training would provide a more demanding assessment of whether the models can transfer to unfamiliar domains. This source distribution may underrepresent contemporary personal and institutional names, modern place references, conversational syntax and journalistic expression. The held-

out test units were sampled from the same combined source pool, not from independently collected modern discourse. The reported F1 values therefore do not establish transfer to spontaneous speech, contemporary community communications or news. Future collection should include community-approved contemporary communications and journalistic materials, with source-held-out and independently sourced out-of-domain evaluations.

The number of examples in two entity categories also limits interpretation. The 95 ANIM annotations are occurrences in the texts, not a complete list of animal terms in Nya-Həba; only 12 such mentions appear in the test set. TIME has 35 test mentions and was also the category with the lowest human entity agreement. With support this small, a few changed predictions can substantially affect category-level F1. An expanded corpus should therefore include more varied examples of temporal and animal expressions, including both familiar terms and forms not previously encountered in the dataset. Small and imbalanced training support may contribute to poor recognition, but the present experiments do not separate that effect from lexical variation, contextual ambiguity or model architecture. To address this, future ANIM collection will sample domestic-animal and wildlife expressions from livestock and agricultural discourse, folktales, oral histories and wildlife descriptions. TIME expansion will sample traditional and contemporary festivals, seasonal activities, cultural ceremonies, calendars and community-event narratives. Annotation will retain the existing context-dependent rules, prioritising distinct lexical forms in natural utterances over repeated isolated word lists.

Muazu (2009) documents phonological and lexical correspondences between the two varieties. Without verified unit-level dialect provenance, a defensible dialect-stratified pilot analysis cannot be reported; this should be addressed through provenance-aware collection and dialect-tagged evaluation in a future release.

5. Conclusion

This study has produced an annotated Nya-Həba corpus for named entity recognition containing 2,038 annotation units and 27,123 tokens. Following adjudication, the corpus contains 1,928 entity mentions across the four categories. Before adjudication, the two fluent-speaker annotators agreed on 98.03% of tokens; Cohen's κ was 0.8631, while exact entity-span and type agreement yielded an F1 of 86.17%. On the common held-out test set, the CRF achieved 75.35% entity F1 and the BiLSTM-CRF achieved 68.19%. Beyond these baseline scores, the work provides the annotation guideline, the original annotations, a record of adjudication decisions and the scripts used in corpus preparation and model evaluation. These materials offer a starting point for further Nya-Həba language technology research. The next stage should broaden the range of text sources and dialect information and compare the present baselines with pretrained multilingual and African-focused models across several random seeds and source-held-out test settings. Future development will prioritise contemporary-domain and source-held-out testing, targeted ANIM and TIME enrichment through community-approved contextual materials, and explicit Hong/Gaya provenance labels before any dialect-stratified evaluation.

Author Contributions

Conceptualization, C.I.; methodology, C.I.; software, C.I.; validation, C.I.; formal analysis, C.I.; investigation, C.I.; resources, C.I.; data curation, C.I.; writing—original draft preparation, C.I.; writing—review and editing, C.I., A.M.Y. and I.J.D.; visualization, C.I.; supervision, A.M.Y. and I.J.D.; project administration, C.I.; funding acquisition, not applicable.

Funding

This research received no external funding.

Data Availability Statement

The annotated Nya-Həba corpus, the two pre-adjudication annotation versions, the corrected CoNLL data, the annotation guideline and associated reproducibility materials are available through a public repository. The corpus and reproducibility materials are available through Zenodo, (2026). Access to and reuse of Bible-derived source passages remain subject to the applicable third-party source-use permissions.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors thank Mr. Ndachimya Edward Magaji for annotation, Gennesis Simon for adjudication and annotation correction, and Richard Dauda (Maksha Gong), Rev. Simon Yohanna and Mrs. Asenah Abama for the language material identified in the provenance account. They acknowledge the Həba Language Development Association and Nya-Həba Bible Translation for their written authorization, and Evang. Yahi Emmanuel Alaji Gaya and Adulea B. Bulama for signing the authority letter on behalf of the respective bodies. This research was not funded by any grant.

References

Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., Mayhew, S., Azime, I. A., Muhammad, S. H., Emezue, C. C., Nakatumba-Nabende, J., Ogayo, P., Anuoluwapo, A., Gitau, C., Mbaye, D., Alabi, J., Yimam, S. M., Gwadabe, T. R., Ezeani, I., Niyongabo, R. A., Mukiibi, J., Otiende, V., Orife, I., David, D., Ngom, S., Adewumi, T., Rayson, P., Adeyemi, M., Muriuki, G., Anebi, E., Chukwuneke, C., Odu, N., Wairagala, E. P., Oyerinde, S., Siro, C., Bateesa, T. S., Oloyede, T., Wambui, Y., Akinode, V., Nabagereka, D., Katusiime, M., Awokoya, A., Mboup, M., Gebreyohannes, D., Tilaye, H., Nwaike, K., Wolde, D., Faye, A., Sibanda, B., Ahia, O., Dossou, B. F. P., Ogueji, K., Diop, T. I., Diallo, A., Akinfaderin, A., Marengereke, T., & Osei, S. (2021). MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics, 9, 1116–1131.* https://aclanthology.org/2021.tacl-1.66/

Adelani, D. I., Neubig, G., Ruder, S., Rijhwani, S., Beukman, M., Palen-Michel, C., Lignos, C., Alabi, J. O., Muhammad, S. H., Nabende, P., Dione, C. M. B., Bukula, A., Mabuya, R., Dossou, B. F. P., Sibanda, B., Buzaaba, H., Mukiibi, J., Kalipe, G., Mbaye, D., Taylor, A., Kabore, F., Emezue, C. C., Aremu, A., Ogayo, P., Gitau, C., Munkoh-Buabeng, E., Memdjokam Koagne, V., Tapo, A. A., Macucwa, T., Marivate, V., Mboning, E., Gwadabe, T., Adewumi, T., Ahia, O., Nakatumba-Nabende, J., Mokono, N. L., Ezeani, I., Chukwuneke, C., Adeyemi, M., Hacheme, G. Q., Abdulmumin, I., Ogundepo, O., Yousuf, O., Moteu Ngoli, T., & Klakow, D. (2022). MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 4488–4508). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.298

Agić, Ž., Hovy, D., & Søgaard, A. (2015). If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 268–272). Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-2044

Alabi, J. O., Adelani, D. I., Mosbach, M., & Klakow, D. (2022). Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 4336–4349). International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.382/

Alabi, J. O., Amponsah-Kaakyire, K., Adelani, D. I., & España-Bonet, C. (2020). Massive vs. curated embeddings for low-resourced languages: The case of Yorùbá and Twi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 2754–2762). European Language Resources Association. https://aclanthology.org/2020.lrec-1.335/

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34(4), 555–596.* https://doi.org/10.1162/coli.07-034-R2

Asiimwe, P. A., Meyer, F., & Buys, J. (2026). RunyaNER: Auxiliary language selection for Runyankore NER. In *Proceedings of the 6th Workshop on Multilingual Representation Learning (MRL 2026)* (pp. 480–491). Association for Computational Linguistics. https://aclanthology.org/2026.mrl-1.38/

Balama, K. H. (2024, October 8). Milestone in promotion of Həba language and culture. News Hunter. https://thenewshunterblog.wordpress.com/2024/10/08/milestone-in-promotion-of-heba-language-and-culture/

Christodoulopoulos, C., & Steedman, M. (2015). A massively parallel corpus: The Bible in 100 languages. *Language Resources and Evaluation, 49, 375–395.* https://doi.org/10.1007/s10579-014-9287-y

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20(1), 37–46.* https://doi.org/10.1177/001316446002000104

Falola, P. B., Alabi, J. O., Akinola, S. O., Ogunajo, F. T., Alabi, E. O., & Adelani, D. I. (2026). YoNER: A new Yorùbá multi-domain named entity recognition dataset. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference* (pp. 7137–7148). https://doi.org/10.63317/4o6m2hzpy7gs

Glottolog. (n.d.). Huba [Language catalogue entry; ISO 639-3: hbb; Glottocode: huba1236]. Retrieved November 23, 2026, from https://glottolog.org/resource/languoid/id/huba1236

Hedderich, M. A., Adelani, D. I., Zhu, D., Alabi, J., Markus, U., & Klakow, D. (2020). Transfer learning and distant supervision for multilingual transformer models: A study on African languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2580–2591). Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-main.204/

Huang, Z., Xu, W., & Yu, K. (2015). *Bidirectional LSTM-CRF models for sequence tagging* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1508.01991

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 282–289). Morgan Kaufmann.

Muazu, M. A. (2009). The Hong and Gaya dialects of Kilba. *California Linguistic Notes, 34(2), 1–7.*

Ogueji, K., Zhu, Y., & Lin, J. (2021). Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning* (pp. 116–126). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.mrl-1.11

Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., & Ji, H. (2017). Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1946–1958). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1178

Pustejovsky, J., & Stubbs, A. (2012). *Natural language annotation for machine learning: A guide to corpus-building for applications.* O'Reilly Media.

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003* (pp. 142–147). Association for Computational Linguistics. https://aclanthology.org/W03-0419/

Wycliffe Bible Translators, Inc., & Nigeria Bible Translation Trust. (2018). *Taɗər Alkawal Aku Nya Həba* [Nya-Həba New Testament]. Nigeria Bible Translation Trust. https://www.bible.com/versions/2723-hbbnt-taɗər-alkawal-aku-nya-həba

Zenodo. (2026). Zenodo Open Research Repository. *https://doi.org/10.5281/zenodo.23328858*