

Counting What Exists: A Repeatable Protocol for Building Language Resources Where None Exist, with a First Case from Nigeria

Oyinbrakemi Praise Oyimi Omare

Twokriks, independent research · praiseoyimi526@gmail.com

Abstract

Speech and language systems now reach users in developing countries at scale, but they reach them in the few languages the systems were built from. The people left out are those who speak only languages with no recorded data, and they are often the people least able to switch. We present a repeatable protocol for building a resource where almost nothing exists: count what exists, credit it, record new speech only with consent, structure every entry the same way, and measure progress against a fixed record. We apply it first to three languages of Nigeria's South South: Nigerian Pidgin (pcm), Izon (ijc) and Urhobo (urh). Open speech models now list all three, but the open speech behind those claims is a few hundred utterances of Izon and none of Urhobo, and no measurement of how well any model hears them has been published. We count the starting state: two Izon dictionaries cleaned to machine readable text (about 8,500 and 10,500 headwords), a Pidgin wordlist of 336 entries, 293 of them glossed by a first language speaker of the Warri variety, and the Naija treebank of 9,241 sentences used as a reference for word selection. Collection has begun. We state the protocol, a first paid round of nine speakers designed to yield 6,000 unique items and about 14 hours of speech per language, a planned first measurement of open speech models on these languages, and the limitations of a resource begun by one glosser from one town. All counts are reproducible from the released files.

1 Introduction

Use of large language models has grown faster in low and middle income countries than in high income ones (Chatterji et al., 2025). A conversation level study of users in four such countries found that Nigerian users had the highest median engagement of the four, 91 conversations per user, and that 96.2% of their task oriented conversations were in English while only 50.6% of their casual conversations were (Roy Chowdhury and Garimella, 2026). The pattern is consistent with a population that uses these

systems for work in a language the systems support and speaks to them in that language only when it has to. The study did not detect which non English languages were used, because its detector separated English from everything else; that limitation is itself part of the picture.

The languages at issue here are spoken mainly in the South South geopolitical zone of Nigeria (Akwa Ibom, Bayelsa, Cross River, Delta, Edo and Rivers states). Nigerian Pidgin is spoken across the country as a first or second language; Izon has an estimated one million speakers (Lewis, 2009); Urhobo is spoken mainly in Delta State. Speech models now list all three, but none is served in the way English, or increasingly Hausa, Yoruba and Igbo, now are. A speaker of Izon who wishes to use these systems must do so in a second language, and older speakers, for whom that second language is weakest, are the least served.

This paper does not present a finished dataset. It presents a method and the starting state of its first application, counted, so that progress can be measured against this record. The method does not depend on the language or the country. Nigerian Pidgin, Izon and Urhobo are the first three because we can reach their speakers; Itsekiri follows, and the same steps apply to any language where the record is this thin. Our contributions are:

1. A collection protocol, independent of language, for growing a resource from a counted starting state, with a first round of 6,000 unique items and about 14 hours of speech per language from three paid speakers each, and consent, compensation and quality checks built in (Section 5).
2. Machine readable text versions of two published Izon dictionaries (Kolokuma and Nembe), with counts of headwords, tone marks and dotted vowels retained after digitisation (Section 4).
3. A Nigerian Pidgin wordlist of 306 forms selected from the Naija treebank by frequency and glossed in the Warri variety, with 30 register items absent from the corpus (Section 4).
4. A plan for the first published measurement of how well open speech models that list these languages actually transcribe them (Section 6).

2 Languages

Nigerian Pidgin (pcm; Naija) is an English lexified creole spoken across Nigeria with regional varieties. The variety of interest here is that of Warri, Delta State, which differs lexically from the Lagos and Port Harcourt varieties recorded in most existing corpora. Orthography is not standardised; the Naija treebank's conventions (Caron et al., 2019) are the most widely used in computational work.

Izon (ijc; Ijaw, Ịzọn) is an Ijoid language of Bayelsa, Delta, Rivers, Ondo and Edo states. Kolokuma is the dialect used in education and the basis of the main published dictionary (Williamson and Blench). Nembe (nmb) is treated as a separate language in Glottolog and has its own dictionary (Kaliai, 1964; edited Blench, 2008). The orthography marks tone and uses the dotted vowels ị, ọ and ụ.

Urhobo (urh) is an Edoid language of Delta State (Ethiopia East, Ughelli, Sapele, Warri). A print dictionary exists (Ukere, 1986). **Itsekiri** (its), spoken in and around Warri, is in scope for a later phase and has no resource in this release.

3 Related work

For Pidgin, the NaijaSynCor project produced a spoken treebank of 9,242 sentences and 140,729 tokens with English translation and audio alignment (Caron et al., 2019), later released in Universal Dependencies as pcm_nsc; it remains the only treebank for the language. Pidgin appears in MasakhaNER 2.0 (Adelani et al., 2022) and MasakhaNEWS (Adelani et al., 2023) through BBC Pidgin news text, and in small speech sets: a 4,277 utterance parallel set (Rufai et al., 2026), the 31 hour NaijaS2ST set, and 1.63 hours of Pidgin accented English in AfriSpeech-200 (Olatunji et al., 2023). Meta's MMS model lists pcm among its ASR languages (Pratap et al., 2024). For Hausa, Igbo and Yoruba, NaijaVoices adds about 1,800 hours under a non commercial licence (Emezue et al., 2025); none of it covers the languages here. None of these corpora targets the Warri variety.

Meta's Omnilingual ASR (Omnilingual ASR Team, 2025), released under the Apache 2.0 licence, lists Nigerian Pidgin, Izon, Urhobo and Itsekiri among more than 1,600 languages, and its open corpus holds 466 Izon and 532 Itsekiri utterances and none in Urhobo or Pidgin. MMS released text to speech models for all four under a non commercial licence (Pratap et al., 2024). A language being listed is not the same as a language being served: we find no published error rate for any of these models on Izon, Urhobo or Warri Pidgin speech. Beyond these, we find no speech corpus in Izon or Urhobo. AfriSpeech-200 contains English speech from speakers whose accent is tagged Ijaw (9.2 hours), Izon (2.7 hours) or Urhobo (1.9 hours), which is useful for accent robust English recognition but is not data

in these languages. Text is limited to scripture and the JW300 corpus; Orife (2020) trained translation baselines for four Edoid languages including Urhobo on JW300, which is the only published NLP result for Urhobo we are aware of. Dictionaries exist in print and as catalogue entries (Williamson and Timitimi, 1983; Kaliai, 1964; Ukere, 1986) but not as machine readable lexicons. Dialect level resource work on a Nigerian language is exemplified by Ahia et al. (2024) for Yorùbá, whose framing of regional varieties we follow.

The difference from each of these is narrow and should be stated: we add machine readable Izon lexicons where only print existed; a Warri variety Pidgin wordlist where corpora cover other varieties; and, once recorded, speech in Izon and Urhobo where almost none is open, with the first measurement of how well the models that list them perform.

4 The resource as it stands

4.1 Sources and cleaning

The Kolokuma dictionary (Williamson and Blench) and the Nembe dictionary (Kaliai, second edition, edited by Blench, 2008) were converted from their circulated PDF form to plain Unicode text. No manual correction has yet been applied. We report the state of the text as it is, including digitisation damage, so that later cleaning can be measured.

The Pidgin wordlist was built from the Naija treebank by token frequency and split into two sheets: forms that are not English words (158 entries) and English forms doing a different grammatical job in Pidgin, such as *go* as a future marker or *make* as a permissive (148 entries). Each entry carries the corpus frequency, the corpus gloss and a gloss in the Warri variety written by the author, a first language speaker of that variety. A third sheet holds 30 words and phrases of the Warri register that do not occur in the treebank at all. Of the 336 glosses, 43 (42 English forms and one register item) were drafted from the author's stated usage rules and await his confirmation; they are marked as such. The treebank supplies the selection of forms only: its sentences are not redistributed, and example sentences in the released list are our own.

4.2 Counts

Resource	Unit	Count	Status
Kolokuma dictionary text	characters / words	1,018,792 / 171,107	released, uncorrected
	headwords (est.)	~8,500	
	combining tone marks	12,565	
	dotted vowels i ɔ u ɛ	31,974	
Nembe dictionary text	characters / words	692,197 / 104,214	released, uncorrected
	headwords (est.)	~10,500	
	combining tone marks	6,392	
	dotted vowels	17,049	
Pidgin wordlist, non English forms	entries / glossed	158 / 158	released
Pidgin wordlist, English forms, new jobs	entries / confirmed	148 / 106	42 drafted, to confirm
Warri register additions	entries / confirmed	30 / 29	1 to confirm
Naija treebank (reference)	sentences / tokens	9,241 / 140,837	external, CC BY-SA; forms only
Izon numerals primer	pages	33	held, not released
Urhobo parables	texts	10	held, not released
Round 1 items, batch 1 (English, our own)	items	1,500	written; batches 2 to 4 to follow
Recorded speech	hours	begun	counted in next version

Table 1: The resource at the time of writing. Headword counts are estimates (Section 7). Treebank figures are our recount over the three released CoNLL-U files and differ by one sentence and 108 tokens from the figures in Caron et al. (2019).

5 Collection protocol

Material is organised as items. An item is a word or a short everyday sentence written in English by the project, never taken from an existing corpus, together with its translation typed and recorded by one speaker of each language. No item repeats: across batches, and between the speakers of a language, every item is different, so the round covers as many words and constructions as its budget allows. Translation runs as a chain: Pidgin speakers translate from English, and Izon and Urhobo speakers translate from the

English and the Pidgin together, so the Pidgin form is fixed before the others are written. Natural speech is collected as talks of about five minutes on everyday topics the project sets, such as a market, a journey or a story heard from a grandparent; talks are the main source of the code switching between English, Pidgin and Izon that real users produce. Nothing is recorded covertly in public places. Background noise from markets is recorded separately, without intelligible speech, for use in training robustness.

The first round runs for two months with three speakers per language, nine in all, each working in four batches of two weeks. In each batch a speaker translates and records 500 items of their own and records eight talks, about an hour of work a day. Per language the round yields 6,000 unique items, about 6 hours of read speech, plus 8 hours of talks: about 14 hours. Across the three languages, about 42 hours. That passes the roughly ten hours per language at which adapting an open speech model begins to produce usable output (Pratap et al., 2024). Later rounds work toward 100 hours of Pidgin and 50 hours each of Izon and Urhobo.

Speakers are adults who grew up speaking the language. Each signs an agreement for their own language before any recording and is paid ₦100,000 for the round, in four payments made only on accepted batches: ₦900,000 for the round, about ₦21,000 (US\$16 at ₦1,330 to the dollar) per hour of speech. Speakers work on a web recorder on their own phone or computer and see only their own work; device and sample rate are logged per file. In the data each speaker is a code with age range, gender, first language, and dialect or town as the speaker names it. Izon dialect is recorded with care, since our two dictionaries cover only Kolokuma and Nembe and speakers may use another variety.

Quality is checked at three points, and every rating is made by the project lead, never by speakers. Every speaker passes a ten item test before the first batch. Each batch hides ten check items that every speaker of the language receives, the only items that repeat, so answers can be compared across speakers and against a key. Automatic checks flag recordings that are silent, clipped or rushed, and the project lead reviews every flag and a sample of each batch. A batch is accepted when at least 95 of every 100 recordings are usable and the audio matches the text; anything short is returned to the speaker to redo.

Glossing of the Pidgin list is done by a single speaker of the Warri variety. We intend to add a second Warri speaker for the English forms sheet and to report agreement between them before claiming the list as validated.

6 Planned baseline

No model has been trained. The first measurement, to be reported in the next version of this paper, is word and character error rate for open speech models that list these languages: Omnilingual ASR at two sizes (Omnilingual ASR Team, 2025) and MMS 1B (Pratap et al., 2024). They are run on a held out test set from our own first round recordings, split so that no test speaker is used in later adaptation, and for every language on read items and talks separately. For Pidgin we also report the share of errors that fall on forms in our Warri wordlist. Character error rate is reported alongside word error rate because tone marks and dotted vowels make word level scoring harsh. These numbers fix the reference point against which an adapted model is later compared, and, as far as we can find, will be the first published for Izon and Urhobo speech and for Warri Pidgin.

7 How the counts were made

Characters and words are straight counts over the cleaned text files. Headwords are estimated as lines beginning with a word form followed by a part of speech abbreviation and are uncertain by several hundred in each direction; the Nembe file yields more headwords than the larger Kolokuma file because its entries are shorter. Tone marks are Unicode combining characters (category Mn); dotted vowels are the characters *ĩ*, *õ*, *ũ* and *ẹ* in either case. Wordlist figures are row counts in the three sheets. Treebank figures are counts of sentence identifiers and token lines over the released train, development and test files. Scripts will be released with the data.

8 Conclusion

We have stated a protocol for building a language resource where almost none exists and counted the starting state of its first application, three languages of Nigeria's South South. The protocol is the part meant to travel: the same steps apply to the next language and the next country. The next version of this paper reports the first recorded batches and the first measured error rates.

Limitations

The Pidgin glosses reflect one speaker of one town's variety and will not generalise to Lagos or Port Harcourt Pidgin; no inter annotator agreement is yet reported. The dictionary texts are uncorrected and tone survived digitisation unevenly, so a share of examples will need a speaker's confirmation before use. Headword counts are estimates. Round 1 uses three speakers per language reached through the project's own network and recording on their own devices, so speaker variety and recording conditions are

narrow and uneven, and the read items reflect one team's choice of topics. Each item is read by one speaker only, which widens coverage but means no sentence is heard in two voices. Forty three Pidgin glosses remain drafts. The resource contains no transcribed speech yet, so nothing here supports a claim about speech recognition for any of the three languages; the coverage claims of existing models are reported as their authors state them and are untested by us until Section 6 is run. Izon coverage is limited to two dialects, and Urhobo text is limited to ten parables. The usage figures cited in the introduction come from a convenience sample of 243 Nigerian users recruited through a gig platform and describe that population, not the country; no regional breakdown within Nigeria is available in any source we found.

Ethical considerations

The dictionaries are the work of their compilers and are redistributed as text for research; the copyright position of each will be stated with the release and the compilers credited. Speakers sign an agreement before any recording, are paid for every accepted batch, and are identified in the data only by a code with age range, gender, first language and dialect. A speaker's voice is used as a voice in a product only with separate, optional consent. The agreement follows the Nigeria Data Protection Act 2023. Nothing is recorded covertly. Recordings will not include personal information beyond the speaker's voice. The resource is intended for speech and language research and for language learning; it is not intended for speaker identification. The work is self funded.

References

- David Ifeoluwa Adelani et al. 2022. MasakhaNER 2.0: Africa centric transfer learning for named entity recognition. In *EMNLP*.
- David Ifeoluwa Adelani et al. 2023. MasakhaNEWS: News topic classification for African languages. In *IJCNLP AACL*.
- Orevaoghene Ahia et al. 2024. Voices unheard: NLP resources and models for Yorùbá regional dialects. arXiv:2406.19564.
- Bernard Caron, Marine Courtin, Kim Gerdes and Sylvain Kahane. 2019. A surface syntactic UD treebank for Naija. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories*.
- Aaron Chatterji, Tom Cunningham, David Deming et al. 2025. How people use ChatGPT. OpenAI, September 15, 2025.
- Chris Emezue et al. 2025. The NaijaVoices dataset: Cultivating large scale, high quality, culturally rich speech data for African languages. In *Interspeech*. arXiv:2505.20564.
- M. H. I. Kalai. 1964. *Nembe English dictionary*, second edition. University of Ibadan Press. Edited version, Roger Blench, 2008.
- M. Paul Lewis, ed. 2009. *Ethnologue: Languages of the World*, 16th edition. SIL International.

- Tobi Olatunji et al. 2023. AfriSpeech-200: Pan African accented speech dataset for clinical and general domain ASR. *TACL* 11.
- Iroro Orife. 2020. Towards neural machine translation for Edoid languages. arXiv:2003.10704.
- Vineel Pratap et al. 2024. Scaling speech technology to 1,000+ languages. *JMLR*.
- Omnilingual ASR Team. 2025. Omnilingual ASR: Open source multilingual speech recognition for 1600+ languages. arXiv:2511.09690.
- Shreyasi Roy Chowdhury and Kiran Garimella. 2026. How people use ChatGPT: Conversation level evidence from India, Nigeria, Brazil, and Pakistan. arXiv:2609.38279.
- Rufai et al. 2026. A Nigerian Pidgin speech corpus. *PMLR* 302.
- A. O. Ukere. 1986. *Urhobo English dictionary*. Ilupeju Press.
- Kay Williamson and Roger Blench. *Dictionary of Kolokuma Nzon*. Circulated draft.
- Kay Williamson and A. O. Timitimi. 1983. *Short Izon English dictionary*. University of Port Harcourt Press.