RAN: Resource Abundance Notation for Languages in NLP

Jared R. Coleman[1] **Tainã G.D. Coleman**[2] **Bhaskar Krishnamachari**[3]
[1]Loyola Marymount University [2]San Diego Supercomputer Center
[3]University of Southern California
jared.coleman@lmu.edu t1coleman@sdsc.edu bkrishna@usc.edu

Abstract

The term "low-resource" is used pervasively in NLP but communicates almost nothing precise. We propose **RAN (Resource Abundance Notation)**, a compact, multi-dimensional notation for quantifying a language's NLP resource profile. A RAN score is written as S/M/L_1-B_1/L_2-B_2/. . . , where $S = \lfloor \log_{10}(\text{speakers}) \rfloor$, $M = \lfloor \log_{10}(\text{monolingual sentences}) \rfloor$, and each L_i-B_i pair records a bilingual partner and $\lfloor \log_{10}(\text{parallel sentences}) \rfloor$. Values derive from canonical sources: Wikidata for speakers, OSCAR 23.01 for monolingual corpora, and (where available) OPUS for parallel corpora. We score 20 typologically diverse languages (including Quechua, Guarani, Cherokee, and Owens Valley Paiute) and correlate each profile against published benchmarks for machine translation (MT, via NLLB-200 chrF++), named entity recognition (NER, via XTREME XLM-R WikiANN F1), and part-of-speech tagging (POS, via XTREME XLM-R UD accuracy). The RAN components carry complementary information: a linear model using all three explains 52% of MT variance, 76% of NER variance, and 72% of POS variance. Among single predictors, $B_{\max}$ (the largest bilingual corpus, regardless of partner) is strongest for the cross-lingual transfer tasks (NER, POS), while M and B_{en} are strongest for MT. RAN is designed first as a *communication* tool, not a predictive model.

1 Introduction

"Low-resource" appears in thousands of NLP papers with no shared definition. A paper may use the term whether a language has 500 parallel sentences or 500,000, whether it has 10 fluent speakers or 10 million. This makes it difficult to compare results across papers, assess whether a technique might transfer to a new language, or communicate the difficulty of a task. Existing classifications, such as Joshi et al. (2020) provide coarse, single-dimensional categories that are hard to reproduce from first principles and too imprecise for quantitative reasoning.

We propose **RAN (Resource Abundance Notation)**, designed first as a *communication* tool. When a paper reports evaluation on a `4/2/en-4` language (Cherokee) versus a `7/4/en-7/fr-6` language (Swahili), a reader instantly grasps the difference across multiple dimensions, without looking up corpus statistics or consulting a classification table. The notation is **compact** (fits in an abstract), **multi-dimensional** (separating speakers, monolingual data, and bilingual data), **reproducible** (derived from canonical, citable sources), and **interpretable** (each integer is an order of magnitude).

This matters especially for Indigenous languages, where the heterogeneity of "low-resource" is extreme: Quechua (`6/0/en-6/es-2`), Guarani (`6/1/en-6/es-2`), Cherokee (`4/2/en-4`), and Owens Valley Paiute (`1/3/en-3`) differ by orders of magnitude in speaker vitality and corpus availability yet are often bundled under a single label.

2 Notation

A RAN score is written as:

$$\text{S/M/}L_1\text{-}B_1\text{/}L_2\text{-}B_2\text{/. . .} \quad (1)$$

where $S = \lfloor \log_{10}(\#\text{ fluent speakers}) \rfloor$, $M = \lfloor \log_{10}(\#\text{ monolingual sentences}) \rfloor$, and $B_i = \lfloor \log_{10}(\#\text{ parallel sentences with language } L_i) \rfloor$, with B_i listed in descending order.

Each dimension corresponds to a different kind of NLP training or downstream use. S measures the *community*: language vitality, the realistic ceiling on future annotation or participatory data work, and the downstream population any tool will serve. It is the only dimension that cannot be grown by scraping, and it motivates the ethical and sovereignty considerations that corpus counts alone do not capture (we expand on this in §5). M measures the raw

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 168–172

material for self-supervised pretraining: the text a self-supervised language model (e.g., BERT, GPT, XLM-R) can consume, and therefore the ceiling on what any monolingual or multilingual encoder can learn about the language without alignment. B_i is the supervised counterpart: parallel sentences are the direct training data for MT, and they enable cross-lingual transfer from a high-resource partner. L_i records the partner identity, because a language paired with English behaves differently from one paired only with a regional neighbor. Listing the pairs in descending B_i order puts the strongest connection ($B_{\max}$) first. This is the quantity most relevant for cross-lingual transfer (§4), and the ordered list itself reads as a pivot map (§8).

We intentionally report M and the B_i as counts in their *source corpora* (OSCAR and OPUS respectively) rather than as a single combined figure. This keeps each component independently citable and reproducible from a canonical, dated snapshot. As a consequence, a configuration like $M = 1$, $B_{\text{en}} = 6$ (e.g. Guarani in our data) is internally consistent: it means OSCAR 23.01 contains $\sim$10 Guarani sentences while OPUS lists $\sim 10^6$ Guarani–English pairs. A reader who needs the *total* monolingual text available, ignoring source provenance and duplication, can read $\max(M, \max_i B_i)$ off the notation directly as a lower bound.

The $\lfloor \log_{10} \rfloor$ quantization is primarily a communication choice, but it also matches what is known about how NLP performance responds to data: cross-entropy loss scales as a power law in corpus size (Kaplan et al., 2020; Bansal et al., 2022), so meaningful differences between languages appear at the order-of-magnitude level rather than in raw counts. The gap between 100 and 1,000 sentences is qualitative, while the gap between 100,000 and 100,900 is invisible. Integer values capture this directly, fit in an abstract, and are comparable across languages at a glance. A useful side-effect is robustness to upstream noise: deduplication policy, sentence-splitter choice, and domain coverage can shift raw counts by tens of percent without changing the floored log. When a task calls for finer precision, a decimal form ($M = 3.4$ for $\sim$2.5K sentences) coexists with the integer notation.

3 Data

We scored 20 typologically diverse languages spanning the full RAN range, from English (`9/10/...`) to Owens Valley Paiute (`1/3/en-3`). **Speaker counts** are the maximum P1098 value across statements on each language's Wikidata entity (Q-ids in `languages.csv`), typically the L1+L2 total. **Monolingual corpus sizes** are estimated from OSCAR 23.01 deduplicated word counts (OSCAR Project, 2023), converted to sentences at 15 words/sentence (5 for logographic scripts). Hausa, absent from OSCAR 23.01, falls back to CC-100. **Bilingual data** comes from OPUS (Tiedemann, 2012), queried with `preprocessing=moses`. For each pair we record the size of the *largest single corpus* OPUS lists, not the sum: this avoids double-counting derived/aliased releases (e.g. NLLB derives from CCMatrix; HPLT/MultiHPLT report identical counts) and corresponds to the realistic ceiling for a single-corpus model. Queries use the ISO 639-1 macrocode (e.g. `ar`, `zh`); sub-variant tags (`arz`, `cmn`, `swh`, etc.) are not aggregated by OPUS and including them would not change any integer bin here. The released `languages.csv` records the OPUS corpus that produced the maximum for every pair, so each value is reproducible. For Owens Valley Paiute (`mnr`), effectively absent from OPUS, both M and B_{en} derive from the community-curated Kubishi Dictionary[1] (4,484 `mnr`–en sentence pairs).

We compare against published benchmark numbers (no models were trained for this work): **machine translation (MT)** via NLLB-200 on FLORES-200 xx→eng, chrF++ (NLLB Team et al., 2024) (17/20 languages); **named entity recognition (NER)** via XLM-R Large zero-shot on WikiANN F1 (Conneau et al., 2020; Hu et al., 2020) (10/20); and **part-of-speech tagging (POS)** via XLM-R Large zero-shot on Universal Dependencies accuracy (9/20). Table 1 gives the full inventory.

4 Correlation with Benchmark Performance

We fit linear regressions predicting benchmark scores from RAN components and compare across a small set of models (Table 2). The linear form is used as a coarse summary of how the integer-log components track benchmark performance. We do not claim that quality is literally linear in $\log(\text{data})$, only that log-scale components capture the order-of-magnitude effects discussed in §2. Our aim is twofold: (a) test whether the three components carry complementary information by comparing

[1] https://dictionary.kubishi.com/

Language	ISO	RAN	sum
English	eng	9/10/es-8/fr-8/de-8/zh-7	27
Spanish	spa	8/9/en-8/fr-8/pt-8	25
Chinese†	zho	9/9/en-7/ja-7	25
Hindi	hin	8/8/en-7/ur-6	23
Arabic†	ara	8/8/en-7/fr-7	23
Vietnamese	vie	7/9/en-7/zh-6	23
Turkish	tur	7/8/en-7/de-7	22
Korean	kor	7/8/en-7/ja-6	22
Sinhala	sin	7/7/en-7	21
Nepali†	nep	7/7/en-7/hi-6	21
Mongolian†	mon	6/7/en-7/zh-5	20
Hausa	hau	7/6/en-6/fr-6	19
Swahili†	swa	7/4/en-7/fr-6	18
Welsh	cym	5/6/en-7	18
Yoruba	yor	7/2/fr-6/en-6	15
Maltese	mlt	5/3/en-7/it-6	15
Guarani†	grn	6/1/en-6/es-2	13
Quechua†	que	6/0/en-6/es-2	12
Cherokee	chr	4/2/en-4	10
OVP‡	mnr	1/3/en-3	7

Table 1: RAN components for the 20 languages, sorted by RAN_{sum}. OVP = Owens Valley Paiute. English–English "bilingual" is reported as $B_{\text{en}} = 0$ by convention. For English we use $B_{\max}$ as its largest partner. † Macro-language ISO 639-3 code (zho, ara, swa, mon, que, nep, grn). The numbers shown are dominated by one variant in practice (e.g. Arabic by Modern Standard *arb*; Quechua by *quy* Ayacucho Southern, the variant used in NLLB-200's `quy_Latn` FLORES split). Sub-variants (e.g. *quz* Cuzco, *qub* Huallaga) would receive distinct RAN scores when scored individually. We recommend reporting the ISO 639-3 code alongside RAN whenever a macro-language label could be ambiguous. ‡ OVP is not in OSCAR or OPUS, so both M and B_{en} derive from the community-maintained Kubishi dictionary (`https://dictionary.kubishi.com/`).

the full model against single-predictor baselines, and (b) identify which single component is the strongest predictor for each task.

Model	MT	NER	POS
$S + M + B_{\max}$	**0.52**	**0.76**	**0.72**
$S + M + B_{\text{en}}$	0.52	0.53	0.47
$B_{\max}$ only	0.41	0.48	0.64
M only	0.26	0.22	0.35
B_{en} only	0.41	0.06	0.12
RAN_{sum}	0.20	0.19	0.37
S only	0.00	0.00	0.05

Table 2: Training R^2 for each predictor model across tasks. MT: NLLB-200 chrF++ (n=17); NER: XTREME WikiANN F1 (n=10); POS: XTREME UD accuracy (n=9). For MT, $B_{\text{en}} = B_{\max}$ for all 17 languages.

Three observations follow from the regression experiment (Figure 1 visualizes the aggregate trend). First, the full three-component model dominates every task: at $R^2 = 0.76$ for NER, no single component is within 0.28, so dropping any of S, M, or $B_{\max}$ leaves real variance on the table. The R^2 gap between the full model and any single component is itself the evidence that the components carry complementary information rather than restating each other. Second, $B_{\max}$ is the strongest single predictor for NER ($R^2 = 0.48$) and POS ($R^2 = 0.64$), while B_{en} alone explains almost nothing on those tasks ($R^2 = 0.06$ and 0.12), confirming that cross-lingual transfer benefits from connections to *any* high-resource partner, not just English. Third, S alone is uninformative of benchmark performance. We argue in Section 5 that this is a feature of the notation, not a defect. Because samples are small ($n \in [9, 17]$), we treat these R^2 values as descriptive rather than as evidence of out-of-sample predictive power.

Chinese sits well below its RAN-predicted score on NER and POS, large enough to ask whether the gap reflects RAN missing something Chinese-specific or a property of logographic, non-segmented scripts more broadly. The other non-Latin-script languages in our set are consistent with the latter reading: Korean (alphabetic syllabary, space-segmented) lands close to its RAN-predicted NER/POS scores, while Vietnamese (Latin, segmented, tonal) shows the same pattern. We therefore attribute Chinese's gap to script and tokenization in the zero-shot benchmark setting (XLM-R transfer from English WikiANN/UD) rather than to a corpus signal RAN fails to capture. Confirming this would benefit from adding a second logographic language (e.g. Japanese) and a non-segmented non-logographic language (e.g. Thai). We leave this to future work.

5 Why Keep Speaker Count?

S's non-predictiveness should not be read as an argument to drop it from the notation. Corpus size tells you what a model *can do today*, while S tells you why we should care and what resources could plausibly become available. It captures language vitality (Swahili `7/4/...` and Maltese `5/3/...` have near-identical corpus profiles but very different endangerment status), the realistic ceiling on future annotation and participatory data work, the scale of downstream impact, and the community capacity and sovereignty considerations that shape whether a dataset should exist at all. A notation

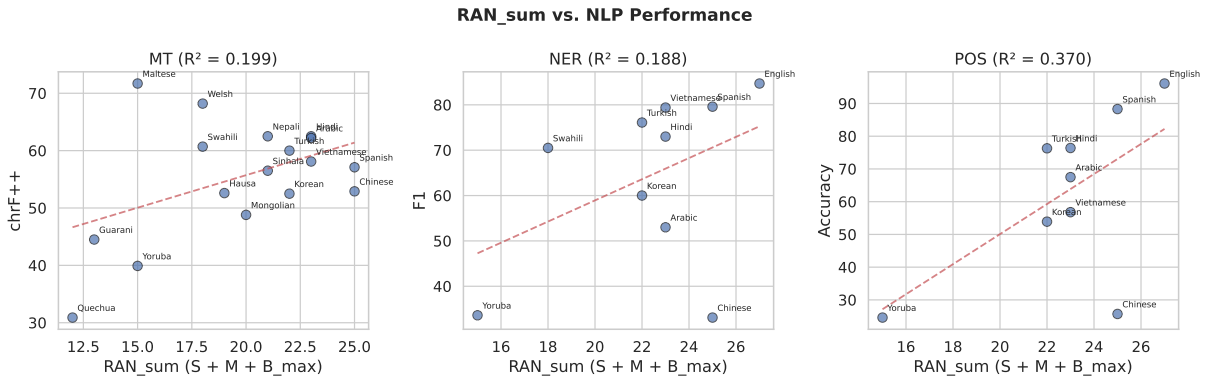

Figure 1: RAN_{sum} vs. benchmark performance across the three tasks, with linear fits.

that omitted S would be a poorer communication tool even if it remained an equally good predictor of benchmark scores.

6 A Living RAN Database

A notation is only as useful as the shared numbers it references. We deploy a community-maintained web database at `https://ran.kubishi.com` where anyone can submit an update (ISO 639-3 code, any subset of S, M, or L_i-B_i pairs with citations). Every submission is queued for human review of source reproducibility and correct denomination (deduplicated sentences, not bytes), and submissions and revisions retain stable IDs so a cited RAN score can always be traced back to the exact revision it was drawn from.

7 Related Work

Joshi et al. (2020) introduced a six-class taxonomy (0–5) for language resource levels, widely adopted but offering only a single dimension. Blasi et al. (2022) demonstrated systematic inequalities in NLP performance across languages, motivating more precise quantification of resource gaps. NLLB (NLLB Team et al., 2024) and XTREME (Hu et al., 2020) provide the benchmark data enabling our empirical validation.

8 Conclusion and Possible Extensions

RAN provides a compact, reproducible, multi-dimensional notation for communicating language resource profiles. Across 20 languages and three tasks, the components carry complementary information ($R^2 = 0.52$ for MT, 0.76 for NER, 0.72 for POS), and $B_{\max}$ outperforms B_{en} for cross-lingual transfer. We frame RAN as a *communication* tool: `6/0/en-6/es-2` (Quechua) and `1/3/en-3` (Owens Valley Paiute) name resource profiles a single "low-resource" label would flatten. We encourage adoption in abstracts, dataset and model cards, and shared-task descriptions (e.g., resource envelopes like "all targets satisfy $M \leq 4$"). Natural extensions of the core notation include rendering the L_i-B_i graph to surface pivot paths, a decimal form for finer precision, and optional dimensions (script status, lexicon size, typological distance, quality flags) appended per task.

This paper deliberately scopes RAN to written text and text-based benchmarks (MT, NER, POS), because the data sources it draws on (OSCAR, OPUS) and the benchmarks it validates against (NLLB-200, XTREME) are themselves text-only. But many of the languages RAN is most useful for (particularly Indigenous languages of the Americas like Owens Valley Paiute, Cherokee, and many Quechuan and Tupian variants) are predominantly or exclusively spoken, and a text-only profile understates their actual resource picture. A natural extension is an optional speech component $H = \lfloor \log_{10}(\text{hours of recorded speech}) \rfloor$, populated from sources such as Common Voice, FLEURS, and community archives, and paired with bilingual H_i values for aligned speech–text resources. This would let RAN describe ASR/TTS resource profiles alongside the text-based one. We leave the design of H and its empirical validation against speech benchmarks to future work, but flag it here as the most important near-term direction for languages of the Americas.

Limitations

Our sample is small ($n \in [9, 17]$ per task) and English-centric: NER and POS benchmarks rely on XLM-R zero-shot transfer from English, and B_{en} coincides with $B_{\max}$ for every MT language.

The floor-of-log formulation compresses very different corpus sizes (e.g. 1K vs. 9K sentences) into the same integer. RAN does not encode data *quality* (domain, noise, script-register match), which can shift effective performance by 10+ chrF++. Counting in *sentences* is also an approximation: corpora differ in average sentence length and complexity, and our words÷15 (or ÷5 for logographic scripts) heuristic only coarsely normalises this. A corpus of short, simple sentences and one of long, syntactically rich sentences with the same M are not equivalent training material. We accept this loss of precision in exchange for a unit that is meaningful across scripts (unlike raw token counts, which are tokenizer-dependent) and that fits in an abstract. The decimal form $M = 3.4$ and per-task quality flags can recover precision where needed. RAN also currently assumes catalogued parallel sentences are human-produced: it does not yet distinguish machine-translated or back-translated pairs, which can now be generated cheaply and would otherwise inflate B_i without a commensurate quality gain, nor does it account for silver/synthetic data more broadly. Speaker counts inherit Wikidata's heterogeneity (different sources, years, L1/L2 conventions). Finally, RAN reports only text resources: until the H extension sketched above is in place, languages with substantial speech corpora but little text will be understated. RAN should complement, not replace, qualitative community-facing context.

Ethical Considerations

Several languages in our dataset (Cherokee, Quechua, Guarani, and Owens Valley Paiute) are spoken by Indigenous communities, some critically endangered. RAN is descriptive and uses only aggregate, publicly cited statistics. It is not intended to rank languages by "worth" or to justify deprioritizing any language. We emphasize that low S (few speakers) is precisely where language-sovereignty considerations and community partnership matter most. Decisions about whether and how to build NLP tools for an Indigenous language must rest with the speaker community, not with corpus counts.

References

Yamini Bansal, Behrooz Ghorbani, Ankush Garg, and 1 others. 2022. Data scaling laws in NMT: The effect of noise and architecture. In *Proceedings of the 39th International Conference on Machine Learning*.

Damian E. Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 5486–5505. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, and 1 others. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. 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*, pages 6282–6293. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, and 1 others. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

OSCAR Project. 2023. OSCAR 23.01. `https://oscar-project.github.io/documentation/versions/oscar-2301/`. Open Super-large Crawled Aggregated coRpus, version 23.01.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218. European Language Resources Association.