

Cross-Lingual Transfer from Portuguese to Nheengatu: Evidence for Contact-Induced Convergence as a Computational Bridge

Rafael Macário Fernandes

University of São Paulo

Department of Linguistics, São Paulo, Brazil

rafael.macario@usp.br

Abstract

Typological distance predicts that Portuguese-to-Nheengatu transfer should fail: the languages are genetically unrelated and typologically divergent (URIEL+ genetic 1.0, syntactic 0.61). We argue that 400 years of contact-induced structural convergence may provide a computational bridge invisible to these measures, so that transfer should succeed at the levels where contact has operated most deeply. Across five experiments on a parallel corpus of about 5,000 sentence pairs, we find a progressive pattern: word-level and lexical-semantic methods fail, while fine-tuned XLM-RoBERTa reaches Precision@1 of 24.7% and MRR of 0.371 on cross-lingual sentence retrieval—more than double a BM25 character- n -gram baseline and $1.7\times$ that of zero-shot LaBSE, the strongest off-the-shelf retriever we test. Crucially, Portuguese beats Spanish—typologically near-identical and no less represented in pretraining, but without comparable contact history—on every metric and all five seeds, and the two sources share substantial lexical overlap, so lexical overlap alone is unlikely to explain the gap. These results suggest that contact history can distinguish transfer sources that typological distance treats as equivalent.

1 Introduction

Nheengatu (ISO 639-3: *yrl*)¹ is a Tupian language and the only living descendant of Tupinambá, the former lingua franca of colonial Brazil (Moore, 2014; g1 Ceará, 2025). According to the 2022 Brazilian census, it is spoken by 13,070 people in Brazil, making it the country’s eighth most spoken Indigenous language (IBGE, 2024). It is co-official

¹Also written *Yēgatu*, or *Yēgatú* with stress marked (Finbow and O’Neill, 2022), following the orthography the Rio Negro communities unified with FOIRN/UFAM in 2016; the language is also known as *Língua Geral Amazônica*. We use the spelling that is standard in the computational literature (Alencar, 2023).

with Portuguese in São Gabriel da Cachoeira (Amazonas) and is also spoken in neighboring Colombia (Moore, 2014; g1 Ceará, 2025). In the Upper Rio Negro, where it serves as a supra-ethnic lingua franca, it has been adopted by several Indigenous groups, while community-led revitalization initiatives are expanding its use in Pará (Cintra, 2018; Silva Meirelles, 2020). Despite its demographic importance and ongoing revitalization, Nheengatu remains severely under-resourced: existing computational research relies on essentially the same small set of textual resources as ours (Adelani et al., 2024; de Alencar, 2024; Pinhanez et al., 2024).

Portuguese is the natural candidate for cross-lingual transfer: every parallel resource in our corpus pairs Nheengatu with it, and Adelani et al. (2024) find that zero-shot POS transfer into Nheengatu is stronger from Brazilian Portuguese than from English (31.5 vs. 28.9 accuracy). Typologically, however, the pair is unfavorable. URIEL+ (Khan et al., 2025) places the two at maximal genetic distance (1.0) and at syntactic distance 0.61, against 0.38 for Spanish–Portuguese.² Since transfer is “substantially less effective . . . for distant languages” (Lauscher et al., 2020) and typological distance generally guides transfer-source selection (Lin et al., 2019; Rice et al., 2025; Ng et al., 2026), transfer for this language pair should be weak.

What these measures miss is 400 years of intense language contact. Portuguese conjunctions have grammaticalized in Nheengatu (adversative *ma*, explanatory *porke*), each more fixed than in the source (Cândido et al., 2021), and Finbow and O’Neill (2022) document pronominal restructuring toward Portuguese, borrowed preverbal modals (*pudéri* <

²Computed with the *urielplus* package (database integration plus imputation). We report URIEL+ rather than URIEL (Littell et al., 2017), which has no observed syntactic features for Nheengatu and returns the same distance for every language with an empty syntax vector. Sparsity is systemic: 87% of URIEL+’s typological matrix is imputed (Ng et al., 2025).

poder ‘can’), and a rapidly expanding GO-future under pressure from Brazilian Portuguese *ir* + infinitive. Alves (2026) finds the same auxiliary constructions in annotated corpus data. In our corpus, hundreds of loanword types—roughly 6% of the vocabulary—include legal borrowings such as *estado* → *istadu* ‘state’ (Cruz, 2011; Avila, 2021). These convergences motivate our central claim:

Contact-induced structural convergence acts as a *de facto* typological bridge that may enable cross-lingual transfer—but not uniformly. For Portuguese and Nheengatu, transfer should be strongest at the linguistic levels where such convergence has been independently documented. Computationally, this predicts a *progressive* pattern: methods operating at shallower levels should fail, whereas methods exploiting supra-lexical representations—those spanning more than the word—should succeed.

We make three contributions: (i) a *contact-informed hierarchy* linking independently documented levels of contact to the computational levels that probe them; (ii) the first evaluation of cross-lingual sentence representations for Nheengatu; and (iii) evidence from a controlled Portuguese–Spanish comparison for a contact-related signal beyond typological similarity. Five experiments follow: four probing successive contact levels, one a source-language control.

2 Related Work

Cross-lingual transfer. Multilingual transformers such as mBERT (Pires et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) enable cross-lingual transfer in low-resource settings (Lauscher et al., 2020). Transfer is strongest when the low-resource target has a structurally similar high-resource partner: Senel et al. (2024) show that using a high-resource cousin language substantially boosts NLU across five low-resource Turkic languages. For Indigenous languages of the Americas, Ebrahimi et al. (2022) show that zero-shot transfer is weak but fine-tuning on small amounts of target text substantially improves results. Adelan et al. (2024) demonstrate that Brazilian Portuguese outperforms English as a transfer source for low-resource Brazilian languages, motivating Portuguese as a plausible candidate. When direct

parallel data is scarce, synthetic training data produced by machine translation (“translate-train”) is a standard control (Conneau et al., 2018b), though it can introduce translationese artifacts (Artetxe et al., 2020).

Contact linguistics and transfer. Contact linguistics provides the theoretical basis for our hypothesis. Thomason and Kaufman (1988, 35) hold that contact outcomes are primarily determined by the sociolinguistic history of speakers rather than by the structure of their languages—so contact can reshape grammar in ways typological distance metrics cannot anticipate. Law (2013) documents contact-induced borrowing of bound morphemes in Lowland Mayan, producing structural convergence unlikely between unrelated languages. For Portuguese–Nheengatu specifically, such convergence at the clause-combining and constructional levels is well documented (Cândido et al., 2021; Finbow and O’Neill, 2022). Alves (2026) measures that convergence: by word order entropy over 24 syntactic features, Nheengatu patterns with Portuguese rather than Tupinambá in subject, object, and root position; in auxiliaries and copulas, which Tupinambá lacks; and in subordinator placement. He quantifies where it is detectable; we test whether it aids transfer. More broadly, cross-lingual transfer has been shown to succeed where source and target languages share deep structural, phonological, and morphological properties (Bankula and Bankula, 2025; Haddad-Najjar and Abu-Rabia, 2024; Barking et al., 2025), and to operate differentially across linguistic levels (Cosma, 2026; Blevins et al., 2022).

Nheengatu NLP. Prior computational work on Nheengatu is limited and largely rule-based or resource-building. de Alencar (2021) implements a Nheengatu grammar fragment in Grammatical Framework with Portuguese and English modules—structural overlap made computationally exploitable, but through hand-crafted rules rather than learned representations. The same line of work yields the Yauti analyzer (Alencar, 2023) and the first Universal Dependencies (UD) treebank for the language (de Alencar, 2024; Alexandre and Alencar, 2026). Alexandre et al. (2021) compile a tagged Nheengatu corpus. Neural work is scarcer: Pinhanez et al. (2024) fine-tune translators from Nheengatu into English, and prior work on Indigenous-language generation has used Spanish

as a pivot with NLLB-200 (Fernandes, 2026). All of this takes the transfer source as given; we ask *which* source transfers, and why. Ours is the first evaluation of cross-lingual sentence representations for Nheengatu, and the first source comparison that holds typological distance constant—Adelani et al. (2024) contrast English with Portuguese, which differ in typology as well as contact history.

3 Corpus

The corpus comprises 4,997 Portuguese–Nheengatu sentence pairs spanning legal, narrative, historical, and pedagogical registers, drawn from four sources: the Brazilian Constitution translated into Nheengatu (2,020 pairs); historical fragments from the Tycho Brahe corpus (Veronesi, 2026) (597 pairs); and Navarro’s pedagogical grammar (Navarro, 2016) (2,139 pairs) and Casasnovas’s grammatical notes (Casasnovas, 2006) (241 pairs), both obtained via the CompLin/nheengatu repository (Alencar, 2023). Quality filtering removed 31 misaligned pairs via length-ratio outlier detection ($\pm 3\sigma$) followed by manual verification. All resources—code, datasets, trained models, and the machine-translated Spanish control set—are publicly available.³ The *hapax rate* (the proportion of word types occurring exactly once) is reported in Table 1. Nheengatu’s hapax rate of 60.6% greatly exceeds Portuguese’s 45.5%, reflecting its productive agglutinative morphology and the resulting lexical sparsity. All four sources are written and edited; no spoken or informal Nheengatu is represented, so the contact features we measure are those that have entered the written standard.

Metric	Portuguese	Nheengatu
Total word tokens	66,990	59,334
Unique word types	6,653	8,149
Hapax rate (freq. = 1)	45.5%	60.6%
Type/token ratio	9.9%	13.7%

Table 1: Vocabulary statistics for the cleaned corpus (4,997 pairs).

Beyond aggregate statistics, the corpus directly attests the contact phenomena our hypothesis concerns (Table 2): phonologically adapted loanwords, the borrowed adversative *ma* (61 tokens), preverbal modal auxiliaries of the *pudéri* family (408 tokens),

³Code: <https://github.com/rmaacario/nheengatu-crosslingual> Data and checkpoints: <https://huggingface.co/rmaacario/nheengatu-xlmr>

the *sú* + Verb construction (37 pairs), and the relativizer *waá* (476 tokens).

Portuguese	Nheengatu
LEXICAL <i>polícia</i> ‘police’	<i>pulisia</i>
CLAUSE-COMBINING ... <i>mas</i> ele não foi ‘... but he did not go’	... <i>ma</i> aé nti usú
CONSTRUCTIONAL: MODAL ... não <i>poderá</i> ser superior... ‘... may not exceed...’	... ti <i>upuderi</i> usasá...
CONSTRUCTIONAL: RELATIVIZER ... cidade grande <i>que</i> está... ‘... a big city that is...’	... tauauasu uiku <i>uaá</i> ...

Table 2: Contact phenomena attested in the corpus, with English glosses. Corresponding items in **bold**. The borrowed modal *upuderi* (< *poder* ‘can’) precedes the verb, where native Nheengatu auxiliaries follow it. The relativizer *uaá/waá* (< Tupi -*ba’e*) occurs with any predicate without nominalizing it, like Portuguese *que*, whereas Tupi–Guaraní subordination nominalizes the verb (Finbow and O’Neill, 2022).

4 Experiments

Task and metrics. Our reference evaluation is *cross-lingual sentence retrieval*: given a Portuguese query sentence, rank all 1,000 Nheengatu sentences in the test set and retrieve the correct translation. *Precision@k* (P@k) measures whether the correct translation appears in the top *k* results; *Mean Reciprocal Rank* (MRR) is the mean reciprocal rank of the first correct retrieval. Retrieval measures whether representations from two languages share a semantic space; it is standard in XTREME via BUCC and Tatoeba (Hu et al., 2020; Artetxe and Schwenk, 2019), and needs no target-language annotation beyond parallel sentences—relevant here, since no NLI or QA benchmark exists for Nheengatu, which AmericasNLI does not cover (Ebrahimi et al., 2022). Experiments 1–4 use the same 80/20 train/test split (3,997 training pairs; 1,000 test pairs). Experiments are grouped by the linguistic level each method probes (Figure 1), mirroring the documented contact-informed hierarchy. *Constructional* names the documented phenomenon—schematic patterns with open slots, such as *sú* + [Verb]; *supra-lexical* names what we measure, since retrieval sees whole sentences and cannot isolate constructions from other signal above the word. (In Construction Grammar words

and morphemes are constructions too, so the morphological level covers bound affixation: *-itá*, *ta-*.)

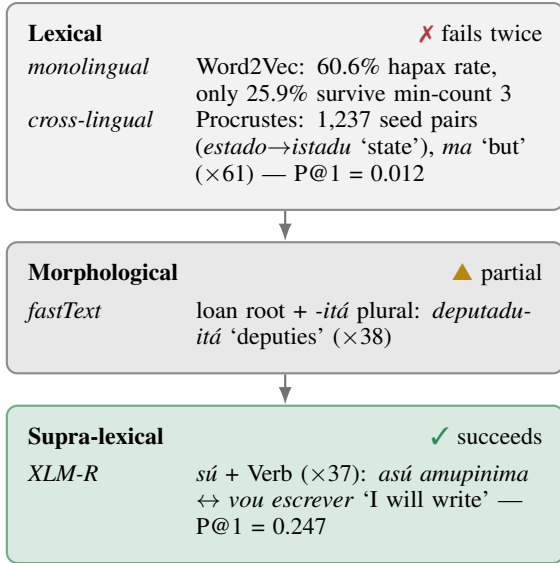


Figure 1: Contact-informed hierarchy: each linguistic level, the methods probing it, and examples attested in our corpus (×*n* = corpus token/pair counts). The lexical level is probed twice—within Nheengatu and across the language pair—and produces only a weak signal (P@1 = 0.012, some 110× the 1/9,048 random baseline), insufficient for retrieval; morphological patterns are captured by fastText but do not scale to sentence retrieval. Strong transfer sufficient for retrieval emerges only at the supra-lexical level.

Experiment 1: Word2Vec. *Goal:* Establish whether word-level semantic spaces are learnable from this corpus. Skip-gram models (100d, window sizes 3 and 10, min_count=3, 10 epochs) were trained separately on each language. At min_count=3, Word2Vec retains only 2,108 Nheengatu types (25.9% of 8,149 unique types); the discarded 74.1% include morphologically productive forms occurring only once or twice. Retained types do not organize semantically: IsoScore 0.023 (Portuguese 0.012), against 0.955 for an isotropic Gaussian of the same dimensions (Rudman et al., 2022). The corpus is therefore too sparse for reliable word-level semantic learning.

Experiment 2: fastText. *Goal:* Test whether character n-grams capture the morphological level of contact. Skip-gram models with character n-grams (minn=3, maxn=6, min_count=1, 20 epochs) (Bojanowski et al., 2017) were trained on each language. The model encodes Nheengatu affixes as consistent directions: for five of eight morphemes documented by Cruz (2011) and Avila (2021),

the derived-minus-root offset is far more coherent within a morpheme than across morphemes (plural *itá*, *d* = 2.10; 3pl *ta-*, *d* = 2.37), and borrowed roots participate: *istadu* → *istaduita* ‘states’, *prupusta* → *prupustaita* ‘proposals’. fastText captures morphological patterns consistent with documented contact phenomena, but mean-pooled fastText sentence vectors achieve P@1 = 0.000, MRR = 0.009: morphological signal does not scale to sentence retrieval.

Experiment 3: Orthogonal Procrustes alignment. *Goal:* Test whether lexical-semantic alignment is achievable given contact-induced lexical overlap. We learn a linear mapping from Portuguese fastText vectors (Common Crawl, 300d) to Nheengatu vectors (trained from scratch, 300d), seeding it with 1,237 automatically extracted Portuguese–Nheengatu candidate correspondence pairs (including many non-loan look-alikes)⁴ (Artetxe et al., 2018). The alignment fails: on 248 held-out seed pairs P@1 = 0.012 and P@5 = 0.028, and even fitted and scored on all 1,237 pairs it reaches only P@1 = 0.039—it cannot recover its own training correspondences.⁵ These rates are some 110× the 1/9,048 random baseline (fastText’s case-sensitive vocabulary), so a weak lexical-semantic signal is present, as the documented borrowing predicts, but far too sparse for retrieval. The from-scratch Nheengatu space clusters by surface form rather than meaning (neighbors dominated by morphological variants), and linear alignment needs semantically organized spaces on both sides.

Experiment 4: XLM-R fine-tuning. *Goal:* Test whether sentence-level representations succeed where word-level methods fail. Sentence encoders can exploit supra-lexical information—sentence embeddings have been shown to encode linguistic properties such as tense and clause coordination (Conneau et al., 2018a), as well as argument frames (Kann et al., 2019)—of the kind restructured through contact (Table 2) (Cândido et al., 2021; Finbow and O’Neill, 2022), without re-

⁴Generated by applying Nheengatu phonological adaptation rules (Cruz, 2011) to Portuguese vocabulary and keeping pairs whose adapted form occurs in our Nheengatu corpus; this high-recall seed is deliberately broader than the validated loanword inventory used in Experiment 5.

⁵We fit on 989 pairs and score 248 held out, so the figures measure generalization, not memorization. Matched capacity does not help: retraining Portuguese from scratch on the same corpus gives P@1 = 0.02 against 0.01 for the pretrained source, under leave-one-out on 100 shared seed pairs.

quiring semantically organized word spaces. We fine-tune XLM-RoBERTa **base** (270M parameters) (Conneau et al., 2020) using Multiple Negatives Ranking Loss (batch size 16, lr 2×10^{-5} , 3 epochs, max length 128, in-batch negatives yielding 15 negatives per anchor, no early stopping), which maps parallel sentence pairs to nearby points in a shared embedding space while separating non-parallel pairs (Figure 2). As baselines we evaluate BM25 over character 3-grams and two multilingual sentence encoders zero-shot—LaBSE (Feng et al., 2022) and paraphrase-multilingual-MiniLM-L12-v2 (Reimers and Gurevych, 2020)—neither of which covers Nheengatu. Table 3 reports the seed-42 run; single runs of this recipe vary, with five seeds under the Experiment 5 harness giving $P@1 = 0.281 \pm 0.012$ on the same test split. We report seed 42 throughout, since the domain and qualitative analyses below derive from that model.

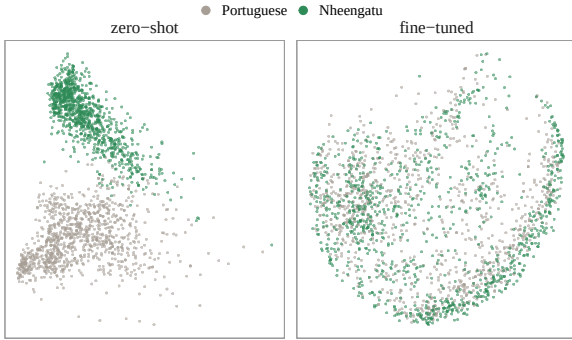


Figure 2: PCA of test-set sentence embeddings ($n = 1,000$ pairs). Before fine-tuning, Portuguese and Nheengatu occupy disjoint regions of the space; after fine-tuning on the training set, they interleave.

System	P@1	P@5	P@10	MRR
Random	0.001	0.005	0.010	0.007
fastText	0.000	0.006	0.013	0.009
XLM-R (z.s.)	0.011	0.042	0.062	0.034
MiniLM (z.s.)	0.036	0.085	0.116	0.066
BM25 (3-gram)	0.104	0.218	0.278	0.165
LaBSE (z.s.)	0.145	0.277	0.346	0.215
XLM-R (ft.)	0.247	0.504	0.621	0.371

Table 3: Sentence retrieval results (test set $n = 1,000$; reference run, seed 42).

Fine-tuned XLM-R outperforms every baseline (Figure 3): more than double BM25 over character 3-grams (0.104) and $1.7\times$ LaBSE (0.145), despite LaBSE’s larger size (471M vs. 270M parameters) and broad multilingual pretraining. Zero-shot XLM-R (0.011) is weaker still, reflecting the

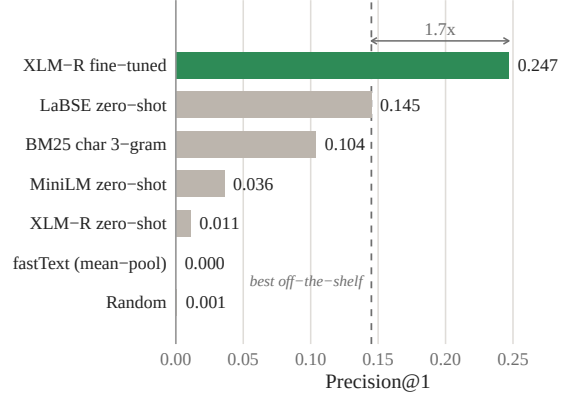


Figure 3: Precision@1 for all retrieval systems (seed 42). Off-the-shelf systems—lexical and neural alike—top out at LaBSE’s 0.145 (dashed); fine-tuning XLM-R on the training set clears that ceiling by $1.7\times$, consistent with the hypothesis that contact history may provide a transfer advantage.

known degeneracy of MLM-only sentence embeddings rather than anything specific to this pair. The advantage is not lexical: on the 799 test queries whose Nheengatu target contains no loanword from our inventory, fine-tuned XLM-R still reaches $P@1 = 0.233$ —close to its 0.247 over the full test set—while BM25 falls to 0.080. And since LaBSE sits only 4.1 points above the character-level baseline where fine-tuned XLM-R sits 14.3, what in-domain fine-tuning adds is not available either from surface matching or from general multilingual pretraining—which is what we would expect of a supra-lexical signal.

Experiment 5: Source-language control. *Goal:* Help isolate contact from generic multilingual transfer. We compare Portuguese against Spanish, which is essentially as distant from Nheengatu as Portuguese is (URIEL+ genetic 1.0 for both; syntactic 0.66 for Spanish against 0.61 for Portuguese, the latter on 183 Grambank features observed for Nheengatu where URIEL has none) and no less represented in XLM-R’s pretraining data (9,374M vs. 8,405M tokens in CC-100; Conneau et al. 2020, Table 6), so pretraining coverage cannot explain a gap, and the two sources are themselves far closer to each other (0.38) than either is to Nheengatu. Following the translate-train paradigm (Conneau et al., 2018b), we machine-translate the Portuguese side of every pair to Spanish with NLLB-200-distilled-600M (Team et al., 2022) and re-run the pipeline,

varying only the source language.⁶ Portuguese wins on every metric and in all five seeds (Table 4); the P@1 gap of $+0.050 \pm 0.007$ is roughly seven times its standard deviation, and at the query level (reference run, seed 42) is significant under approximate randomization for P@1 ($p = 0.00063$) and paired bootstrap for MRR ($p < 10^{-5}$).⁷ Lexical forms are substantially shared: for 56% of the 133 adapted loanwords the Spanish form is string-identical to the Portuguese (*dekretu* < *decreto*, unchanged in Spanish), and mean character similarity to the Nheengatu form is 0.746 for Portuguese against 0.705 for Spanish. Lexical overlap alone therefore does not account for the gap; and because Spanish is derived from the Portuguese source, any translationese (Artetxe et al., 2020) biases it toward Portuguese, making the gap conservative.

Metric	PT (contact)	ES (control)	Gap
P@1	0.281 ± 0.012	0.232 ± 0.005	+0.050
P@5	0.545 ± 0.034	0.453 ± 0.009	+0.091
P@10	0.673 ± 0.030	0.568 ± 0.006	+0.104
MRR	0.406 ± 0.018	0.340 ± 0.006	+0.066

Table 4: Source-language control (mean $\pm$ std over 5 seeds; XLM-R fine-tuning, shared fixed test split, $n = 1,000$). Spanish is machine-translated from the Portuguese source, so the gap is a conservative estimate of any contact-related advantage.

Domain analysis. Performance varies by register (Table 5). The Tycho historical corpus achieves the highest P@1 (0.387). Sentence length does not explain this—the Constitution’s sentences are longer (19.0 vs. 12.3 words) yet score lower—and the two are comparably distinctive against the 1,000 candidates. What separates Tycho is an unusually tight correspondence between the two sides’ token counts ($r = 0.96$ vs. 0.86 for the Constitution): a property of its translation regime, which favors retrieval and bounds how far this figure generalizes. Casasnovas performs lowest (P@1 = 0.057); we report it for completeness ($n = 35$; not interpretable).

⁶Experiments 4 and 5 share the fixed 80/20 split. Table 3 is a single run from our initial implementation, Table 4 five-seed means from the control harness; absolute scores differ between the two, but the paired PT–ES comparison is unaffected, since both conditions share each run’s randomness.

⁷A paired t -test over the five seeds also favors Portuguese ($t = 15.7$, $p < 0.001$), but at $n = 5$ Wilcoxon and the sign test both bottom out at $p = 0.0625$, their minimum achievable value, so we test at the query level instead.

Domain	P@1	Test pairs
Tycho (historical)	0.387	124
Constitution (legal)	0.252	401
Navarro (pedagogical)	0.218	440
Casasnovas [†]	0.057	35
Weighted average	0.247	1,000

Table 5: XLM-R fine-tuned performance by domain.

[†] $n = 35$; not statistically reliable.

Qualitative analysis: the existential construction. Table 6 groups the ten test sentences containing Portuguese *há* by their Nheengatu rendering. The nine *aikué* existentials reach a median rank of 4, and confusions can stay within the construction: for one query, the top-ranked candidate is another *aikué* existential. The lexical exception *Akuera*, by contrast, falls to rank 137. This contrast is consistent with the model representing the existential construction rather than simply matching lexical items, although the sample is too small for firm conclusions.

‘há’ rendered as	Ranks of correct translation
Existential <i>aikué/Aikwé</i> ($n = 9$)	1, 1, 2, 2, 4, 11, 21, 52, 70 (median 4)
e.g.	<i>Aikué igara igarapé upé</i> ‘There is a canoe in the creek’
Lexical exception <i>Akuera</i> ‘long ago’ ($n = 1$)	137
e.g.	<i>Akuera iaeuera aé</i> ‘He has long been used to it’

Table 6: Retrieval ranks (among 1,000 candidates) for the ten test sentences containing Portuguese *há*, grouped by how Nheengatu renders it. The model seems to track the *aikué* construction (median rank 4); the sentence where *há* is not existential falls to rank 137.

5 Discussion

The progressive pattern matches the hypothesis. Word2Vec and Procrustes fail because a 60.6% hapax rate leaves the from-scratch Nheengatu space too anisotropic to organize semantically—the fitted map cannot retrieve even its own training pairs—and fastText’s morphological signal does not scale to sentence retrieval. Sparsity is part of that explanation, but not of the source-language gap: the Nheengatu side is the same text in both conditions, and on the source side Spanish is if anything slightly less sparse than Portuguese (43.4%

vs. 45.5% hapax over the same 4,997 pairs), so sparsity would favor the control—yet Portuguese wins. Fine-tuning succeeds at the supra-lexical level, aligning sentence representations without needing organized word spaces. The Portuguese advantage (Table 4) points to long-term contact rather than generic Romance typology or model capacity, and the domain results agree—register-varied sentences (Tycho, Constitution) outperform the short, formulaic ones (Navarro), while BM25 and LaBSE bound the lexical-overlap contribution. $P@1 = 0.247$ is modest, but retrieval among 1,000 candidates is far harder than bilingual classification. For practical use $P@10 = 0.621$ matters more: the correct translation is in the top ten for nearly two-thirds of queries, enough for translation-memory lookup or parallel-text alignment, though not for unattended retrieval.

6 Conclusion

Our results highlight a factor standard typological resources do not encode: *language contact*—a gap URIEL+’s own maintainers note (Shipton et al., 2026). Two findings support this. First, transfer from a Romance source to Nheengatu succeeds well above lexical and zero-shot baselines despite maximal genetic distance, so typological distance alone under-predicts transfer. Second, of two comparably distant sources, the one with a documented contact history—Portuguese—transfers significantly better, across all metrics and five seeds. The progressive pattern across linguistic levels further suggests that what multilingual sentence encoders exploit is supra-lexical, consistent with contact-induced convergence though not localized to it by our design. We do not claim contact is necessary for transfer, nor that this generalizes beyond this pair and task, but it is a measurable factor in transfer-source quality, worth encoding alongside genealogy and typology.

Future work. Future work will (i) expand the source-language comparison with non-contact and genealogically related (Tupi–Guaraní) controls, complementing Alves (2026); (ii) evaluate additional tasks, including morphological tagging and dependency parsing on the UD_Nheengatu-CompLin treebank (de Alencar, 2024) against the baselines of Alexandre and Alencar (2026), and a Nheengatu natural language inference set following AmericasNLI (Ebrahimi et al., 2022); and (iii) collect and compare spoken Nheengatu against the

written registers studied here.

Ethics Statement

This research uses publicly available written texts and is motivated by supporting Nheengatu language technology for revitalization. Deployment of any models should be conducted in consultation with Nheengatu-speaking communities.

References

- David Ifeoluwa Adelani, A. Seza Doğruöz, André Coneglian, and Atul Kr. Ojha. 2024. [Comparing LLM prompting with cross-lingual transfer performance on indigenous and low-resource Brazilian languages](#). In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 34–41, Mexico City, Mexico. Association for Computational Linguistics.
- Leonel Figueiredo de Alencar. 2021. [Uma gramática computacional de um fragmento do nheengatu / a computational grammar for a fragment of nheengatu](#). *Revista de Estudos da Linguagem*, 29(3):1717–1777.
- Leonel Figueiredo de Alencar. 2024. [A Universal Dependencies treebank for nheengatu](#). In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, pages 37–54, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Leonel Figueiredo de Alencar. 2023. [Yauti: A tool for morphosyntactic analysis of nheengatu within the Universal Dependencies framework](#). In *Proceedings of the 14th Brazilian Symposium in Information and Human Language Technology*, pages 144–154, Belo Horizonte, Brazil. Association for Computational Linguistics.
- Dominick Alexandre, Juliana Gurgel, and Leonel Araripe. 2021. [Compilação de um corpus etiquetado da língua geral amazônica](#). In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 427–431, Porto Alegre, RS, Brasil. SBC.
- Dominick Maia Alexandre and Leonel Figueiredo de Alencar. 2026. [Parsing nheengatu: Performance gains for a Brazilian indigenous Universal Dependencies treebank](#). In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 2*, pages 210–219, Salvador, Brazil. Association for Computational Linguistics.
- Diego Alves. 2026. [Measuring the effects of language contact using word order entropy: the case of nheengatu](#). *Linguistics Vanguard*.

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. [A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. [Translation artifacts in cross-lingual transfer learning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.
- Mikel Artetxe and Holger Schwenk. 2019. [Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond](#). *Transactions of the Association for Computational Linguistics*, 7:597–610.
- Marcel Twardowsky Avila. 2021. *Proposta de dicionário nheengatu-português*. Tese (doutorado), Universidade de São Paulo, São Paulo. Departamento de Letras Modernas, Área de concentração: Estudos da Tradução. Orientador: Eduardo de Almeida Navarro. <https://teses.usp.br/teses/disponiveis/8/8160/tde-10012022-201925/pt-br.html>.
- Ajitesh Bankula and Praney Bankula. 2025. [Cross-linguistic transfer in multilingual nlp: The role of language families and morphology](#).
- M. Barking, Maria Mos, and Ad Backus. 2025. [Investigating language transfer from a usage-based perspective](#). *International Journal of Bilingualism*, 29(2):415–438.
- Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. [Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models](#). *arXiv preprint*. ArXiv:2205.11758 [cs.CL].
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. 2017. [Enriching word vectors with subword information](#). *Transactions of the Association for Computational Linguistics*, 5:135–146.
- A. Casasnovas. 2006. *Noções de língua geral ou Nheengatú: gramática, lendas e vocabulário*. Faculdade Salesiana Dom Bosco.
- Maria Sílvia Cintra. 2018. [Nheengatu language and its role in the tactics of construction of indigenous identity](#). *Advances in Anthropology*, 8(4):289–301.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 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.
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018a. [What you can cram into a single \$\\$&\!#^*\$ vector: Probing sentence embeddings for linguistic properties](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Adrian Cosma. 2026. [An in-vitro study on cross-lingual generalization in language models](#). *arXiv preprint*. ArXiv:2605.26683 [cs.CL].
- Aline da Cruz. 2011. *Fonologia e Gramática do Nheengatú: A língua geral falada pelos povos Baré, Warekena e Baniwa*. LOT Dissertation Series. LOT, Utrecht. https://www.lotpublications.nl/Documents/280_fulltext.pdf.
- Gláucia Vieira Cândido, Aline da Cruz, and Giovana Alves de Oliveira. 2021. [Empréstimos de conjunções e subordinadores do português em nheengatu](#). *Estudos Linguísticos (São Paulo. 1978)*, 50(1):91–107.
- Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. [AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ariavazhagan, and Wei Wang. 2022. [Language-agnostic BERT sentence embedding](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.
- Rafael Fernandes. 2026. [USP at AmericasNLP 2026 shared task: Culturally-aware image captioning for indigenous languages via vision-language models and fine-tuned neural machine translation](#). In *Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 264–271, San Diego, California, USA. Association for Computational Linguistics.
- Tom Finbow and Paul O’Neill. 2022. [Koinéization and language contact: The social causes of morphological](#)

- change in and with portuguese. In Adam Ledgeway, John Charles Smith, and Nigel Vincent, editors, *Periphrasis and Inflexion in Diachrony: A View from Romance*. Oxford University Press.
- g1 Ceará. 2025. Vídeo: Conheça o nheengatu, uma das línguas indígenas mais faladas no brasil. <https://g1.globo.com/ce/ceara/noticia/2025/10/24/video-conheca-o-nheengatu-uma-das-linguas-indigenas-mais-faladas-no-brasil.ghtml>. Accessed: 2026-08-01.
- Laila Haddad-Najjar and Salim Abu-Rabia. 2024. Exploring cross-linguistic transfer of phonology, morphology, syntax, and orthography from Arabic (L1) to English (FL): An intervention study. *International Journal of Bilingualism*, 29(6):1657–1680.
- 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.
- IBGE. 2024. *Censo Demográfico 2022: Indígenas: principais características das pessoas e dos domicílios, por situação urbana ou rural do domicílio: resultados do universo*. Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro.
- Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 287–297, New York City, New York. ArXiv:1811.10773 [cs.CL].
- Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2025. URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6952, Abu Dhabi, UAE. Association for Computational Linguistics.
- Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.
- Danny Law. 2013. Inherited similarity and contact-induced change in mayan languages. *Journal of Language Contact*, 6(2):271 – 299.
- Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.
- Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.
- Denny Moore. 2014. Historical development of nheengatu (língua geral amazônica). In Salikoko S. Mufwene, editor, *Iberian Imperialism and Language Evolution in Latin America*. University of Chicago Press.
- Eduardo de Almeida Navarro. 2016. *Curso de Língua Geral (Nheengatu ou Tupi Moderno): A língua das origens da civilização amazônica*, 2 edition. Centro Angel Rama, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo. 2.ª edição corrigida e aperfeiçoada.
- York Hay Ng, Phuong Hanh Hoang, and En-Shiun Annie Lee. 2025. Less is more: The effectiveness of compact typological language representations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25805–25816, Suzhou, China. Association for Computational Linguistics.
- York Hay Ng, Aditya Khan, Xiang Lu, Matteo Saloum, Michael Zhou, Phuong Hanh Hoang, A. Seza Doğruöz, and En-Shiun Annie Lee. 2026. Modality matching matters: Calibrating language distances for cross-lingual transfer in URIEL+. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 110–130, Rabat, Morocco. Association for Computational Linguistics.
- Claudio Pinhanez, Paulo Cavalin, and Julio Nogima. 2024. Human evaluation of the usefulness of fine-tuned English translators for the Guarani mbya and nheengatu indigenous languages. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, pages 32–36, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

- Enora Rice, Ali Marashian, Hannah Haynie, Katharina von der Wense, and Alexis Palmer. 2025. [Untangling the influence of typology, data, and model architecture on ranking transfer languages for cross-lingual POS tagging](#). In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 22–31, Albuquerque, New Mexico. Association for Computational Linguistics.
- William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2022. [IsoScore: Measuring the uniformity of embedding space utilization](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3325–3339, Dublin, Ireland. Association for Computational Linguistics.
- Lütfi Kerem Senel, Benedikt Ebing, Konul Baghirova, Hinrich Schuetze, and Goran Glavaš. 2024. [Kardeş-NLU: Transfer to low-resource languages with the help of a high-resource cousin – a benchmark and evaluation for Turkic languages](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1672–1688, St. Julian’s, Malta. Association for Computational Linguistics.
- Mason Shipton, York Hay Ng, Aditya Khan, Phuong H. Hoang, Xiang Lu, A. Seza Doğruöz, and Annie En-Shiun Lee. 2026. [Simple additions, substantial gains: Expanding scripts, languages, and lineage coverage in URIEL+](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 11045–11059, Palma de Mallorca, Spain. ELRA Language Resource Association.
- Sâmela Ramos da Silva Meirelles. 2020. *A reinscrição de uma língua destituída: o Nheengatu no Baixo Tapajós*. Tese (doutorado em linguística), Universidade Estadual de Campinas, Campinas, SP.
- 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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. [No language left behind: Scaling human-centered machine translation](#).
- Sarah Grey Thomason and Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. Anthropology: Linguistics. University of California Press, Berkeley.
- Luiz Henrique Lima Veronesi. 2026. *A Plataforma Tycho Brahe: um sistema para corpora sintaticamente anotados*. Tese (doutorado em linguística), Universidade Estadual de Campinas, Instituto de Estudos da Linguagem, Campinas. 211 f.