TRANSFER RUNS THROUGH THE LEXICON: SENTENCE EMBEDDINGS FOR LANGUAGES WITHOUT BITEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual sentence encoders are trained on parallel text, which most of the languages of the world do not have. Nevertheless, many of these languages do have dictionaries. We study a sentence encoder whose only language-specific component is a lexicon, which can be built by hand with native speakers or collected directly from dictionary entries. A student Transformer is distilled from a pretrained sentence embedding model (SONAR) on English text alone. It represents words using their frequency plus a static word embedding derived from a multilingual concept space. We build that space by propagating ConceptNet vectors over the PanLex translation graph, extended with dictionary entries harvested by an LLM. A new language is added by swapping these static word embeddings via lookup tables, with no training. For languages whose lexicon covers at least 85% of running words it reaches 99.1% xsim and 53.7% xSIM++. On all 216 FLORES+ devtest languages, many of them with small lexicons, the model still reaches 60.8% xsim and 23.9% xSIM++, against 21.9% and 10.1% for a training-free baseline that averages the same concept vectors. A key result of this work is an empirical *coverage law*: two statistics available before any training explain R^2=0.85 of the cross-language variance in xSIM++. Deleting known words at random, an intervention that holds the language, the sentences and the encoder fixed, reproduces about four fifths of the slope of the law. Coverage also keeps a partial correlation of 0.90 with quality after controlling for relatedness to high-resource languages (0.30 for relatedness). Finally, we confirm that xsim is saturated for this scenario: it overstates xSIM++ by 36.9 points on average, and it barely registers a 12.8-point xSIM++ difference between two encoders built to differ.

1 INTRODUCTION

Sentence encoders such as LASER (Artetxe & Schwenk, 2019a), LaBSE (Feng et al., 2022) and SONAR (Duquenne et al., 2023) map sentences from many languages into one shared space. That space supports bitext mining (automatic harvesting of parallel translation pairs from monolingual sources) at web scale (Schwenk et al., 2021), which in turn feeds machine translation for low-resource languages (Costa-jussà et al., 2024). These encoders are themselves trained on parallel text. The tool that would bootstrap bitext for a language therefore needs bitext to be built. For most of the roughly 7,000 living languages (Eberhard et al., 2025) no such data exists. Documentation work in these languages usually produces another resource: word lists and dictionaries. The PanLex translation graph (Kamholz et al., 2014) alone links words in more than 6,000 languages.

This paper asks how far a lexicon can replace bitext. We use a strong embedding space (a SONAR teacher) and train a student encoder on English only. The student never reads the input as subwords. Instead, it represents each input word as a static word embedding extracted from a custom multilingual *concept space*, in which translations of a word sit close together. All cross-lingual knowledge then lives in the input layer, which is a lookup table. To embed a sentence in a new language, we swap this lookup table and change nothing else. The lexicon is then the only channel for transfer, so its effect can be measured and manipulated directly.

We evaluate the model on every FLORES+ language and use this setting to answer three questions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- **How far does a lexicon get?** For languages whose lexicon covers at least 85% of running words it reaches 99.1% xsim and 53.7% xSIM++.[1] Across the 216 FLORES+ devtest languages the student scores 60.8% xsim and 23.9% xSIM++. A training-free baseline that averages the same concept vectors scores 21.9% and 10.1%, and the student is ahead of it by more than half a point on 214 of the 216 languages. The SONAR teacher, although not directly comparable because it was trained on massive bitext, remains far ahead (78.8% xSIM++). On languages outside its training set, however, the quality of the teacher drops as their vocabulary overlap with its training languages shrinks: for the six FLORES+ languages with the least overlap it drops to 26.2% xSIM++. Our lexicons for these six are thin, and the student is weaker still (1.2% xSIM++; §5.1).
- **When does it work, and why?** Two statistics available before any training, vocabulary coverage and PanLex density (the size of the PanLex dictionary of a language), predict per-language xSIM++ with R^2=0.85. We call this relationship the coverage law. Randomly deleting known words inside a language reproduces about four fifths of the cross-language slope of the law, and relatedness to high-resource languages explains little once coverage is controlled for. Transfer thus appears to run mainly through the lexicon (§5.2).
- **Is the benchmark trustworthy?** We confirm that xsim, the standard metric for this task, is saturated in this setting. It overstates xSIM++ by 36.9 points for our student. On two encoders built to differ in one known respect, xsim separates them by 0.5 points and xSIM++ by 12.8 (§5.3).

Our contributions are: (i) a novel lexicon-swappable sentence encoder architecture capable of embedding sentences in new languages without bitexts; (ii) the coverage law, supported by observational, interventional and confound-controlled evidence; (iii) a calibration of xsim against xSIM++, which confirms in this setting how much xsim overstates quality; and (iv) ablations of the lexicon and training choices. A single propagation pass over PanLex provides most of the lexicon's gain, and normalizing the distillation target makes training several times faster (§6).

2 Related Work

Multilingual sentence encoders. From recurrent encoders (Artetxe & Schwenk, 2019a) to contrastive Transformers (Feng et al., 2022; Wang et al., 2024) and SONAR and its successors (Duquenne et al., 2023; Omnilingual SONAR Team et al., 2026; Janeiro et al., 2026), these encoders need parallel supervision at scale. Distillation extends an encoder to new languages (Reimers & Gurevych, 2020), and LASER3 trains a student per language (Heffernan et al., 2022), but both align the student with bitext. Our student is distilled on English alone.

Vocabulary and lexicon-based transfer. Artetxe et al. (2020) transfer a monolingual Transformer by retraining its embedding layer on the target language, which is our closest architectural precedent. WECHSEL (Minixhofer et al., 2022) initializes target embeddings from aligned static vectors before further training; Wang et al. (2022) synthesize task data from lexicons; Mullov et al. (2024) decouple the vocabulary of an NMT system; Gurgurov et al. (2024) inject graph-derived adapters. Each of these learns target-side parameters from target-side data. To our knowledge, deriving the input representations of a new language from a lexicon graph and using them by direct substitution, with no training, has not been studied for sentence encoders.

Lexicons, polysemy and LLM harvesting. Bilingual lexicon induction treats the dictionary as a target or a fixed seed (Glavaš & Vulić, 2020) and is fragile for distant languages (Søgaard et al., 2018; Vulić et al., 2019). Retrofitting averages graph neighbours (Speer et al., 2017). We instead learn to aggregate the set of candidate meanings of a word with a set encoder (Zaheer et al., 2017; Lee et al., 2019). LLMs can exploit lexical resources for low-resource languages: word lists carry most of the value of grammar books (Tanzer et al., 2023; Aycock et al., 2024), and dictionary entries help prompting (Merx et al., 2024) and data generation (Yong et al., 2024).

[1] For simplicity, throughout this paper we refer to xsim and xSIM++ as the percentage of correct matches using the xsim method (Artetxe & Schwenk, 2019b). xsim is computed over the Flores-200 benchmark (Goyal et al., 2022; Costa-jussà et al., 2024), and xSIM++ is computed over its synthetic extension with hard distractors (Chen et al., 2023).

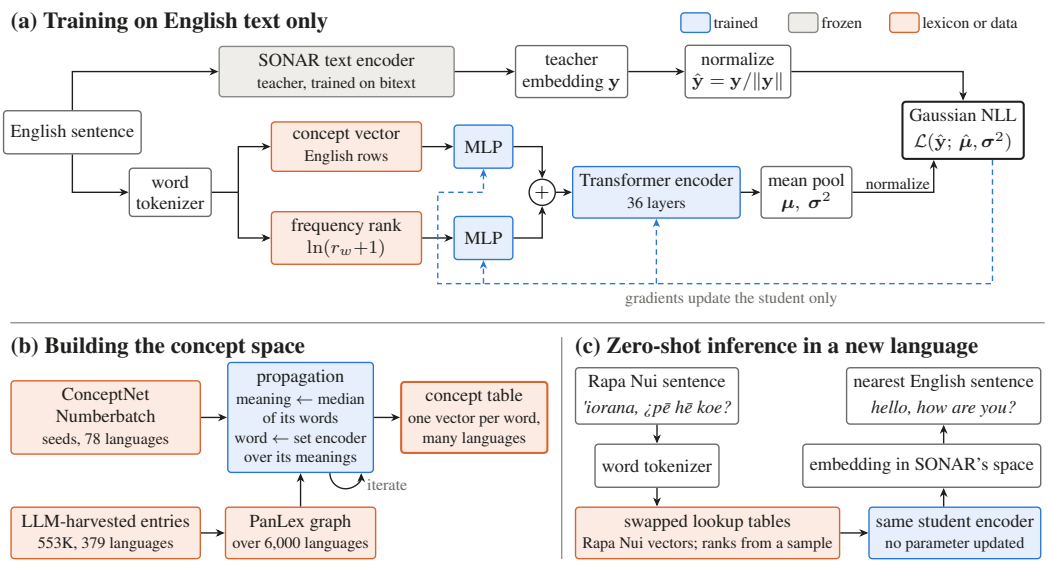

Figure 1: Overview. (a) The student is distilled from the frozen SONAR teacher on English text only, and the concept vectors stay fixed. (b) Numberbatch seed vectors are propagated over PanLex, extended with LLM-harvested entries. (c) Zero-shot inference: only the lookup tables are swapped.

Evaluation and transfer predictors. FLORES-derived benchmarks have quality problems for low-resource languages (Taguchi et al., 2025), and xsim is easy to saturate. xSIM++ counters this with hard distractors (Chen et al., 2023). Studies of what drives cross-lingual transfer point to vocabulary overlap and relatedness (Pires et al., 2019; Wu & Dredze, 2019; K et al., 2020; Conneau et al., 2020), and zero-shot performance of multilingual Transformers can be predicted from language similarity and pretraining data size (Lauscher et al., 2020). In pretrained multilingual models these factors are entangled with pretraining data; in our setting the lexicon is the only channel.

3 A Lexicon-Swappable Sentence Encoder

Model in Figure 1: A student Transformer encoder trained from scratch to reproduce frozen SONAR sentence embeddings from English text only. All cross-lingual capabilities live in the input layer.

Input layer. In contrast to subword tokenization, which fragments many of the words of low-resource languages (Mager et al., 2022), we split texts into whole words. A key reason is that whole words are the unit in which dictionaries are written. Each word w is embedded as the sum of two vectors. The *concept embedding* takes the vector of w in a shared multilingual concept space (below), normalizes it to unit norm, and projects it with an MLP into the encoder's input space. The *frequency embedding* takes the scalar $\ln(\min(r_w{+}1, M))$, where r_w is the frequency rank of w in a monolingual text sample of the language and $M{=}30$K, and projects it with a second MLP. Words absent from the concept space are embedded using a learned unknown vector [UNK], which is shared for new languages. Similarly, punctuation, symbols and short alphanumeric codes, which dictionaries usually do not list, also share a table of 9,448 learned embeddings. During training we add Gaussian noise ($\sigma{=}0.05$) to concept vectors and uniform noise (± 0.5) to log ranks, so that the encoder does not rely on exact values that another language might not reproduce exactly.

Concept space. The concept space is built in four incremental stages, each one attempting to find word vectors for the words that earlier stages could not find. *Seeds*: ConceptNet Numberbatch (Speer et al., 2017) provides aligned 300-dimensional vectors for words in 78 languages. *Propagation*: PanLex (Kamholz et al., 2014) links words and translations across thousands of languages, but its edges do not label word sense. Instead, PanLex connects a polysemous word to several meaning nodes that are shared across languages. Averaging meaning vectors collapses distinct senses into

one point that might not represent either meaning.[2] We therefore compute a vector for every PanLex meaning (the median of the known vectors of its members) and train a permutation-invariant set encoder (Lee et al., 2019) that maps the set of meaning vectors of a word into one single vector. It is trained to reproduce held-out Numberbatch vectors. Applying it to every PanLex word yields vectors for words that have no seed. The procedure can be iterated, since new word vectors change the meaning vectors. *LLM harvesting*: for words that PanLex does not cover, an LLM extracts translation word pairs from dictionaries and word lists collected per language, and a second LLM pass verifies each candidate. The verified entries (553,651 entries in 379 languages, from 643 documents, in our runs) are linked to PanLex meanings and merged into the graph before propagation. This stage is expensive to scale, since many dictionaries exist only as image scans. *Retrieval from unstructured material*: we also test a simpler pipeline built on a retrieval-grounded LLM. The LLM is given several reference documents in the language[3] and is asked for disambiguated glosses in several high-resource languages (for example `eng/human_being`, `spa/persona`). The word receives the vector of its first gloss found in the space. Appendix C gives the prompts, filters and counts.

Multilingual Sentence Encoder and objective. The sentence encoder, which we call the "student", is a Transformer (Vaswani et al., 2017) encoder with mean pooling and a custom head that predicts a mean $\boldsymbol{\mu}$ and a per-dimension variance $\boldsymbol{\sigma}^2$. It is trained on English sentences with the Gaussian negative log-likelihood loss (Nix & Weigend, 1994) of the teacher embedding $\mathbf{y}$:

$$\mathcal{L} = \frac{1}{2} \sum_j \left(\log \sigma_j^2 + \frac{(\hat{y}_j - \hat{\mu}_j)^2}{\sigma_j^2} \right), \qquad \hat{\mathbf{y}} = \mathbf{y}/\|\mathbf{y}\|_2,\ \hat{\boldsymbol{\mu}} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|_2. \tag{1}$$

We drop the input layer of the encoder, a standard learnable embedding layer. In replacement, we directly feed the frozen word embeddings taken from the concept space. In the output layer, we remove magnitude from the regression by normalizing both target and prediction to unit norm (Wang et al., 2017; Steck et al., 2024), due to the fact that retrieval depends only on cosine similarity. The variance head lets the model down-weight inputs it cannot predict confidently, such as sentences with many unknown words (Kendall & Gal, 2017). Section 6.2 compares this objective with plain regression, and measures the effect of the added normalizations.

Zero-shot inference. To embed text of an unseen language, we look up its words in the concept space, estimate frequency ranks from a monolingual text sample, and run the frozen student encoder. Words that are absent from the concept space are replaced by the [UNK] token, and words without a frequency rank are treated as if $r_w \geq M$, so every word has a frequency embedding. No parameter is updated, and no text in the language is needed beyond the sample used for ranks (§6.3).

4 Experimental Setup

Sentence Encoder Model. Implemented using the M2M100 codebase (Fan et al., 2021), with 36 layers and d=1024. All results (except ablations) use one checkpoint, trained from scratch on English NLLB sentences (Costa-jussà et al., 2024) and selected by held-out validation loss (Appendix A). The teacher is the frozen SONAR text encoder. The concept space results from five propagation iterations over the PanLex graph extended with the harvested entries.

Data and metrics. We evaluate English-anchored sentence retrieval on FLORES+ (Goyal et al., 2022): 216 devtest and 224 dev languages, with 1,012 and 997 sentences per language. For each target sentence the model must retrieve its English translation via embedding similarity. Our headline metric is xSIM++ (Chen et al., 2023), which adds hard synthetic distractors to the candidate pool. We also report xsim (Artetxe & Schwenk, 2019b) for comparability. Headline numbers are devtest; dev results are in Appendix D. Words are split with a word-level tokenizer for 164 devtest languages. For

[2] For example, PanLex lists the Yanda Dom Dogon word <*na:*> with two distant meanings, *leg* and *cattle*. The plain average still puts the result far from both of them. Our set encoder places the word embedding closer to one meaning (*cattle*). This encoder learns where to place each embedding informed on reconstruction performance.

[3] This stage uses any reference document found for the language. We include it because it scales more easily, but is not guaranteed to rely only on dictionary-like sources. That is why we measure its impact directly, in Table 3.

Table 1: Retrieval accuracy (%) on FLORES+ devtest, macro-averaged over languages, overall and by the vocabulary coverage of the student (number of languages in parentheses). The two baselines read the same concept vectors as the student, but do not use an encoder. The SONAR teacher is trained on bitext and is shown as a skyline, not as a directly comparable model.

	all (216)		**cov. ≥85% (33)**		**50–85% (105)**		**<50% (78)**	
model	xsim	xSIM++	xsim	xSIM++	xsim	xSIM++	xsim	xSIM++
mean of concept vectors	21.9	10.1	57.6	27.8	24.1	10.6	3.9	1.8
frequency-weighted	26.2	13.9	66.3	37.3	29.8	15.3	4.2	2.1
student (ours)	**60.8**	**23.9**	**99.1**	**53.7**	**79.1**	**28.8**	**20.0**	**4.8**
SONAR teacher (skyline)	97.8	78.8	100.0	88.8	98.6	79.4	95.7	73.8

the other 52, most of them in non-Latin scripts, we use rule-based fallback tokenizers (Appendix B). *Vocabulary coverage* is the share of non-[UNK] running words in a text of a given language. It needs no labels and can be computed on any monolingual text sample. Frequency ranks for the main results are counted independently for each language, on about 2,000 monolingual sentences. §6.3 studies the effect of this sample size.

Baselines. (i) The *SONAR teacher*, a skyline trained on massive bitext. Its tokenizer has language codes for 189 of the 216 languages. For the other 27 we use an "unknown" language tag. (ii) The *mean of concept vectors*: a sentence embedding strategy that consists of averaging the concept vectors the student reads, with no encoder, computed in the same way for the English side and for the target language; plus (iii) a variant weighted by inverse frequency. Baselines (ii) and (iii) use the same lexicon as the student, so they allow us to separate the contribution from the encoder and the concept space. Methods that adapt a model with target-language text (Artetxe et al., 2020; Minixhofer et al., 2022) or bitext (Reimers & Gurevych, 2020; Heffernan et al., 2022) need data that our setting assumes absent, so we do not compare against them.

5 Results

5.1 How far does a lexicon get?

Table 1 shows the main results. First, the encoder adds a large amount to the concept space. The student is 38.9 xsim and 13.9 xSIM++ points above the mean of concept vectors. It is ahead by more than half a point on 214 of the 216 languages in xsim and on 188 in xSIM++, and no language is behind by that margin. The gain is largest at intermediate coverage (79.1 against 24.1 xsim), where averaging word vectors discards most of the sentence while the encoder can still recover it.

Second, quality depends strongly on coverage. When the lexicon covers at least 85% of running words, the student retrieves 99.1% of sentences in xsim. Above 70% coverage, 90 languages average 92.2% xsim. Below 50% coverage, retrieval mostly fails (20.0% xsim). The 52 languages that need the fallback tokenizer all sit in this last group and average 8.6% xsim, against 77.4% for the other 164. Much of the gap between the average and the high-coverage result is therefore a lexicon and tokenization problem, which Section 5.2 quantifies.

Third, the teacher remains far ahead, as expected from a model trained on massive bitext. Of the 27 languages its tokenizer has no code for, 12 are closely related SONAR languages under another code (e.g. `cmn_Hans` for `zho_Hans`) and two are SONAR languages in a new script. On the other 13, the teacher scores 43.5% xSIM++, and its performance correlates with the share of their words that also occur in the training languages of SONAR (Spearman ρ=0.81 for xSIM++ and 0.86 for xsim, reported in Appendix D). For the six with the least overlap, the xSIM++ of the teacher drops to 26.2%. Our lexicons for these six are thin (under 23% coverage), and the student scores only 1.2% xSIM++ on them. The coverage law below puts the coverage it would need to match the teacher at 40% to 81%, depending on the language.

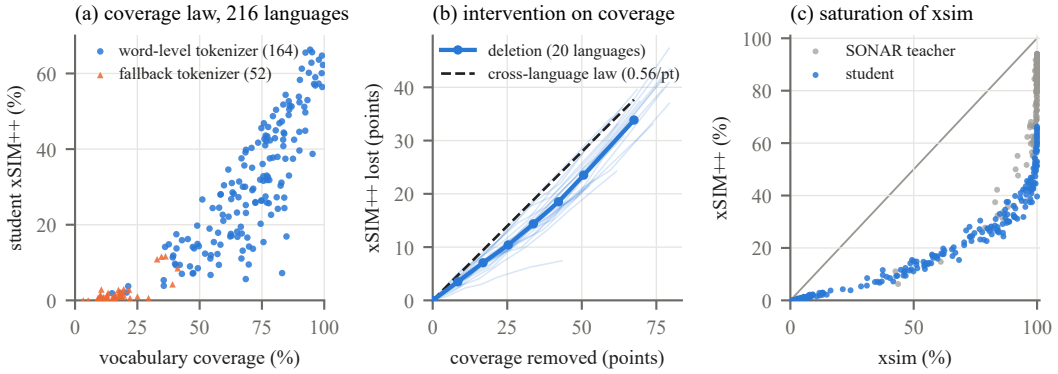

Figure 2: (a) Student xSIM++ vs vocabulary coverage (216 FLORES+ devtest languages). (b) Random deletion of known words in 20 high-coverage languages (thin lines: languages; thick line: mean over languages and three seeds) against the slope of the cross-language regression in (a) (dashed). (c) xsim vs xSIM++ per language. The diagonal marks equal scores.

5.2 The coverage law

In this section we show that the performance of the student is predictable by statistics that are available before any training. This is especially useful to aid practitioners in the decision of relying on the zero-shot performance of the teacher vs the student. As shown below, the former relies heavily on the presence of shared words between the target language and the training languages of SONAR, and therefore performance is capped by a variable out of their control. The latter, instead, is controllable via directed word collection efforts and can increase its xSIM++ with every collected word.

Observational evidence. We regress the xSIM++ of each language on two statistics that are available before any training: *vocabulary coverage* and the *PanLex density* of the language. PanLex density is measured by the logarithm of the number of PanLex entries that link one of its words to a shared meaning (roughly, the size of its PanLex dictionary). The two explain R^2=0.85 of the variance across the 216 languages, and the predicted and measured scores have Spearman ρ=0.94 (Figure 2a). Coverage alone explains 0.79. The relationship does not come from the fallback-tokenized languages clustered near the origin: restricted to the other 164 languages, the two statistics explain 0.78. Each point of coverage is worth 0.56 xSIM++ points; refitting the regression on languages resampled with replacement puts this slope between 0.52 and 0.59 in 95% of the fits. For a new language, a lexicon and a monolingual sample are enough to compute coverage and estimate the expected performance before any training.

Interventional evidence. A cross-language regression cannot separate coverage from what co-varies with it: script, typology, lexicon quality and amount of online text. We therefore change coverage only. In 20 high-coverage languages (mean coverage 84.4%) we replace a random fraction of known words by the [UNK] token at inference, with three seeds. The language, the sentences, and the encoder stay fixed. Figure 2b shows the result. Accuracy falls smoothly, from 47.4% xSIM++ with no deletion to 37.0% at 30% and 28.9% at 50% (93.9, 88.3 and 80.2% in xsim), and the standard deviation across seeds stays at or below half a point. Within a language, each point of coverage removed costs 0.44 xSIM++ points. We estimate this slope from deletion rates up to 50%; resampling the 20 languages gives a 95% interval of 0.41 to 0.46. The slope is about four fifths of the cross-language one. Random deletion does not remove words the way a real lexicon misses them (real gaps concentrate on rare words), so the two slopes are not expected to match exactly. Still, most of the relationship in Figure 2a survives an intervention that touches coverage alone.

Controlling for relatedness. High scores for, say, Aranese Occitan could come from shared vocabulary with well-resourced Romance neighbours rather than from the concept space (Bafna et al., 2023). For each language we compute its genealogical and geographic proximity to the nearest of ten high-resource languages, using URIEL features (Littell et al., 2017) (205 of the 216 languages have entries). Coverage keeps a partial correlation of 0.90 with xSIM++ when relatedness is controlled

Table 2: Calibrating xsim (FLORES+ devtest, 189 SONAR langs.). Each pair differs only in word-order information: the frozen teacher with ordered vs shuffled English input, and two subword English students (Appendix E) with and without positional embeddings. **cos** is the mean cosine similarity between the embedding of each English sentence and the teacher embedding of its translation.

	English side	**xsim**	**xSIM++**	**cos**
frozen teacher	ordered input	99.00	82.26	0.691
	shuffled input	97.66 $\downarrow (-1.34)$	31.07 $\downarrow (-51.20)$	0.414
student	positions on	99.03	83.26	0.682
	positions off	98.54 $\downarrow (-0.50)$	70.43 $\downarrow (-12.82)$	0.543

for. Relatedness keeps 0.30 when coverage is controlled for. In a joint regression on standardized variables the coefficients are 0.88 for coverage and 0.14 for relatedness (both $p < 10^{-4}$). Relatedness has a real effect, but a much smaller one.

5.3 Confirmation of benchmark saturation in this setup

xsim asks whether the true translation is the nearest candidate in a pool of FLORES+ sentences, but most of those sentences are easy to tell apart. xSIM++ adds hard distractors made by editing the true translation (Chen et al., 2023). For our student the two metrics differ by 36.9 points on average (Figure 2c). The gap is smallest below 50% coverage (15.2 points), where both metrics are low, and largest between 50% and 85% (50.3 points). Under xsim, 31 languages appear saturated at 99% or more, while their xSIM++ averages 55.8%. The teacher shows the same pattern at a higher level, with a 16.7-point gap on the languages it supports.

Table 2 compares encoders whose difference is controlled. Shuffling the English input of the frozen teacher costs 1.3 xsim points and 51.2 xSIM++ points, while the sentence vector moves substantially. Training two students that differ only in positional embeddings (enabled vs disabled) gives the same picture: xsim separates them by 0.5 points and xSIM++ by 12.8. Per language, xsim moves by less than half a point for 164 of the 189 languages, while xSIM++ falls by more than five points for 188 of them. We therefore use xSIM++ as the headline metric, and we suggest that future methods that compare with us use that metric.

5.4 Decoding the embeddings

Because the student reproduces the space of SONAR, its embeddings can be fed into the frozen SONAR decoder. When FLORES+ devtest is translated into English for 29 languages chosen to span scripts, families and coverage levels, the student reaches, via greedy decoding, 25.3 chrF++ against 57.3 for teacher embeddings (44%), and chrF++ follows coverage (Spearman ρ=0.97). BLEU is low (3.5 against 33.5), so the outputs are not usable translations. The embeddings carry lexical content that a generator can read, but the decoder quality obeys the same coverage law as retrieval. For the intended use, bitext mining followed by conventional MT training, retrieval quality remains the relevant measure. (Appendix D).

6 What Matters in the Design: Ablation study

6.1 Building the concept space

Table 3 shows the effect of swapping the concept space under the same encoder. Numberbatch alone, which covers only 78 languages, already reaches 41.9% xsim, because many FLORES+ languages are among them or share vocabulary with them. Propagating vectors over the PanLex graph once adds 18.8 xsim and 9.7 xSIM++ points and raises coverage by 11.9 points; 163 languages improve and 9 get worse. The gain arrives with the first iteration. Four further iterations add about 2.7 million vectors to the space, but they change FLORES+ coverage by 0.3 points and scores by less than 0.1. The dictionary entries harvested into PanLex before propagation account for 1.2 points of coverage, 1.7 xsim and 0.7 xSIM++ points. In 48 of the 216 languages it accounts for more than half an xsim

Table 3: Lexicon ablation with the encoder held fixed (FLORES+ devtest, all 216 languages). **cov.** is mean vocabulary coverage; the last column is xsim on the 105 languages with 50–85% coverage under the full lexicon.

concept space	cov.	xsim	xSIM++	xsim, 50–85% coverage
Numberbatch seeds only	45.5	41.9	14.2	46.8
+ propagation over PanLex, 1 iteration	57.3	60.7	23.9	79.0
+ 4 further iterations (full)	57.6	60.8	23.9	79.1
full without dictionary-harvest words	56.5	59.1	23.2	76.3
full without final LLM stage	54.0	57.6	22.5	74.6

Table 4: Training ablations: English-only students without the learned-symbol table, scored every 5K steps on the FLORES+ dev split of 13 validation languages (%). xsim and xSIM++ are measured over the same checkpoints (**step** column), selected by best xsim over the training checkpoints.

positions	loss	L_2 norm	frequency	xsim	xSIM++	step
on	NLL	on	on	57.3	22.4	275K
on	NLL	off	on	58.2	19.9	205K
on	MSE	on	on	59.6	22.7	880K
on	MSE	off	on	36.1	8.0	380K
off	NLL	on	on	53.5	21.3	270K
off	MSE	on	on	54.5	21.7	650K
on	MSE	on	off	60.1	22.3	250K

point and no language reports a negative effect. Seven benefit between 16 and 27 points. The final, less restricted LLM harvest stage adds 3.7 points of coverage, 3.2 xsim points and 1.5 xSIM++ points. It helps 117 languages and hurts none, and it helps most at intermediate coverage.

6.2 Training choices: objective, normalization and word order

Table 4 reports controlled training runs with the same architecture as our model. Each run had a fixed wall-clock budget, so runs stopped at different steps and we compare them at matched steps.

Normalizing the target speeds up training. With the Gaussian NLL, the normalized run reached the best xSIM++ of the unnormalized one (19.9, at 205K steps) after 60K steps, 3.4 times sooner. With MSE the effect is larger: the normalized run reached the best xsim of the unnormalized one (36.1) after 30K steps instead of 380K, 13 times sooner. The unnormalized MSE run was still improving when stopped, so we cannot say whether it would reach a better optimum. At this speed it was not practical, so all other runs used normalization.

The two losses end at similar levels, but NLL gets there sooner. The best xSIM++ is 22.4 for the NLL and 22.7 for MSE with positions on, and 21.3 against 21.7 without. Per step the NLL learns faster: after 10K steps it reaches 45.1% xsim against 5.9% for MSE, and MSE needs about twice as many steps to reach the same level (54.6% at 100K against 54.8% for the NLL at 50K). The NLL runs were trained in float32 because they diverged under bfloat16, and they completed about half as many steps per hour, so the advantage in wall-clock time is smaller.

Word order matters little for this student. Disabling positional embeddings costs about one xSIM++ point under either loss (22.4 to 21.3 for NLL, and 22.7 to 21.7 for MSE). [4]

Masking known words in training helps. Continuing the MSE run from step 880K with 15% of known words replaced by [UNK] beats the same continuation without masking at all 17 matched checkpoints up to 970K, by 2.1 xsim and 1.4 xSIM++ points on average (4.0 xsim on the five validation languages without a devtest split).

[4]Table 2 reports 12.8 points for a different student, one that reads the own subword tokens of SONAR and matches the teacher (83.3% xSIM++). For it, word order is a large share of what remains. Our student reaches about 22%, and we attribute most of its errors to the words themselves (missing or imprecise concept vectors).

6.3 THE FREQUENCY INPUT

The frequency embedding is the only input that needs text in the target language. We test it in two ways. First, we remove it. At step 195K, the last step both runs reached, a student trained without it scores 59.2% xsim and 21.3% xSIM++, against 56.6% and 19.7% for an otherwise identical run with it, both scored by the same evaluator. It is ahead at every scored step after initialization and, at 195K, in all 13 validation languages in xsim. By step 180K it is close to the level the run with frequency reaches only after 880K steps (Table 4). A student without the input would need nothing from a new language except its lexicon. Second, we keep it and vary the text used to estimate frequency ranks. Our main results count ranks on about 2,000 sentences. Counting them on 1,000 sentences gives 60.0% xsim and 22.9% xSIM++ on all 216 languages, against 60.8% and 23.9%. On 20 languages, reducing the sample to 1,000 sentences costs 0.6 xsim and 2.0 xSIM++ points, to 500 sentences 1.3 and 3.4, and to 250 sentences 2.3 and 5.3. A model trained with the input therefore relies on it at inference.

7 DISCUSSION

Scope. We do not aim for the strongest possible encoder for low-resource languages. We use the student to measure how much a lexicon alone can carry, in a setting where the lexicon is the only channel for transfer. All results are therefore conditional on the lexicons we could assemble, and many of them are thin. Because quality follows coverage closely, we expect larger and cleaner lexicons to improve these results with the encoder unchanged.

When to use the student. When bitext exists, training on it remains the stronger approach. Languages that share much of their vocabulary with the training languages of SONAR already receive high teacher scores, which leaves little room for the student. The student targets the remaining languages: those with no bitext and little lexical overlap with any supported language. We expect this group to include many of the world's languages, although FLORES+ contains only a few of them. For such a language, both options can be assessed before any training. The quality of the teacher tracks lexical overlap with its training languages (§5.1), a fixed property of the language. The quality of the student tracks coverage, which grows with every word added to the lexicon. The coverage law turns this into a concrete target. Since a word adds to coverage in proportion to its frequency, collection can start with the most frequent missing words.

What coverage does not explain. At 85% coverage or more, both models saturate xsim, but the student remains 35 xSIM++ points below the teacher (53.7 vs 88.8). Part of this gap likely comes from strings outside the lexicon. xSIM++ distractors often replace entities in the true translation (Chen et al., 2023). The student reads such entities as the shared [UNK] vector, so two sentences that differ only there look alike to it. Representing out-of-lexicon strings, for example by their characters, may recover part of the gap. Segmentation is a second bottleneck: the 52 languages that need the fallback tokenizer average 8.6% xsim, against 77.4% for the other 164. Much of their low coverage may be a segmentation effect, which better word splitting could reduce without new lexical work.

Limitations. Only 13 FLORES+ languages are really outside the language list of SONAR, and our lexicons for most of them are thin, so the comparison with the teacher on unseen languages rests on few languages. Quality on languages without test sets is estimated through the coverage law, which is fitted on FLORES+, so it is yet unclear whether it extrapolates to very different scripts or lexicon sizes. The rule-based fallback tokenizer is weak for scripts without word spacing, and whole-word lookup suits richly agglutinative and polysynthetic languages poorly (Mager et al., 2022).

8 CONCLUSION

A sentence encoder distilled on English alone can reach languages it has never seen through a swappable lexical input layer. How well it does is predictable from the lexicon: vocabulary coverage explains most of the variance across 216 languages, and changing coverage within a language reproduces most of that relationship. Lexicon curation therefore has a measurable effect on how well languages without bitext can be embedded.

AI USE STATEMENT

Large language models are part of the method: LLMs extract and verify dictionary entries, and an LLM-based RAG maps missing words to multilingual glosses (§3; Appendix C describes the prompts). In preparing this work we also used generative AI tools, including coding agents, to assist in some parts of the code implementations, to discover supporting literature, and to draft and edit parts of the paper. All numbers in the paper were produced by the released code from filed experiment outputs and were checked by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Lexical data for minoritized languages encodes community knowledge, and several communities have explicit preferences about its reuse (Ousidhoum et al., 2024). Released lexicons will carry their source provenance. Because the student can embed text in languages with little digital presence, it could support surveillance or scraping of small communities as well as their inclusion; we release it for research on bitext mining and language documentation.

REPRODUCIBILITY STATEMENT

Appendix A gives the architecture of the checkpoint used throughout and all training settings; Appendix B the tokenizer; Appendix C the lexicon pipeline with prompts and counts; and Appendix D per-group and dev results. The training, evaluation and analysis code will be open-sourced, alongside the per-language results for every run, and the scripts that produce every table and figure from the paper.

REFERENCES

Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019a. doi: 10.1162/tacl_a_00288. URL https://aclanthology.org/Q19-1038/.

Mikel Artetxe and Holger Schwenk. Margin-based parallel corpus mining with multilingual sentence embeddings. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3197–3203, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1309. URL https://aclanthology.org/P19-1309/.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. URL https://aclanthology.org/2020.acl-main.421.

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. Can LLMs really learn to translate a low-resource language from one grammar book?, 2024. URL https://arxiv.org/abs/2409.19151.

Niyati Bafna, Cristina España-Bonet, Josef Van Genabith, Benoît Sagot, and Rachel Bawden. Cross-lingual strategies for low-resource language modeling: A study on five indic dialects. In *Actes de CORIA-TALN 2023*, 2023. URL https://aclanthology.org/2023.jeptalnrecital-long.3.

Mingda Chen, Kevin Heffernan, Onur Çelebi, Alexandre Mourachko, and Holger Schwenk. xSIM++: An improved proxy to bitext mining performance for low-resource languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 101–109, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.10. URL https://aclanthology.org/2023.acl-short.10/.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. URL https://aclanthology.org/2020.acl-main.536.

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, Jeff Wang, and NLLB Team. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846, 6 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07335-x. URL https://doi.org/10.1038/s41586-024-07335-x.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot. SONAR: sentence-level multimodal and language-agnostic representations, 2023. URL https://arxiv.org/abs/2308.11466.

D. M. Eberhard, G. F. Simons, and Ch. D. Fennig. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 26th edition, 2025. URL http://www.ethnologue.com.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1), January 2021. ISSN 1532-4435.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.62. URL https://aclanthology.org/2022.acl-long.62/.

Goran Glavaš and Ivan Vulić. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. URL https://aclanthology.org/2020.acl-main.675.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 2022. URL https://aclanthology.org/2022.tacl-1.30.

Daniil Gurgurov, Mareike Hartmann, and Simon Ostermann. Adapting multilingual LLMs to low-resource languages with knowledge graphs via adapters, 2024. URL https://arxiv.org/abs/2407.01406.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2101–2112, 2022.

João Maria Janeiro, Pere-Lluís Huguet Cabot, Ioannis Tsiamas, Artyom Kozhevnikov, David Dale, Belen Alastruey, Alexandre Mourachko, Holger Schwenk, Loic Barrault, Kevin Heffernan, and Paul-Ambroise Duquenne. ECHO: Where multilingual sentence embeddings speak the same language, 2026. URL https://openreview.net/forum?id=nXH1DopZb5.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual BERT: An empirical study. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1912.07840.

David Kamholz, Jonathan Pool, and Susan Colowick. PanLex: Building a resource for panlingual lexical translation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3145–3150, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL https://aclanthology.org/L14-1023/.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision?, 2017. URL https://arxiv.org/abs/1703.04977. NeurIPS 2017.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 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)*, pp. 4483–4499, 2020.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019. URL https://arxiv.org/abs/1810.00825.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 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*, pp. 8–14, 2017.

Manuel Mager, Arturo Oncevay, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. BPE vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022. URL https://arxiv.org/abs/2203.08954.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. Low-resource machine translation through retrieval-augmented LLM prompting: A study on the mambai language. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI)*, 2024. URL https://aclanthology.org/2024.eurali-1.1.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3992–4006, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.293. URL https://aclanthology.org/2022.naacl-main.293/.

Carlos Mullov, Ngoc-Quan Pham, and Alexander Waibel. Decoupled vocabulary learning enables zero-shot translation from unseen languages, 2024. URL https://arxiv.org/abs/2408.02290. ACL 2024.

David A. Nix and Andreas S. Weigend. Estimating the mean and variance of the target probability distribution. *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, 1:55–60 vol.1, 1994. URL https://api.semanticscholar.org/CorpusID:117583961.

Omnilingual SONAR Team, João Maria Janeiro, Pere-Lluís Huguet Cabot, Ioannis Tsiamas, Yen Meng, Vivek Iyer, Guillem Ramírez, Loic Barrault, Belen Alastruey, Yu-An Chung, Marta R. Costa-Jussa, David Dale, Kevin Heffernan, Jaehyeong Jo, Artyom Kozhevnikov, Alexandre Mourachko, Christophe Ropers, Holger Schwenk, and Paul-Ambroise Duquenne. Omnilingual sonar: Cross-lingual and cross-modal sentence embeddings bridging massively multilingual text and speech, 2026. URL https://arxiv.org/abs/2603.16606.

Nedjma Ousidhoum, Meriem Beloucif, and Saif M. Mohammad. Building better: Avoiding pitfalls in developing language resources when data is scarce, 2024. URL https://arxiv.org/abs/2410.12691.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. URL https://aclanthology.org/P19-1493.

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. URL https://aclanthology.org/2020.emnlp-main.365.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1351–1361, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.115. URL https://aclanthology.org/2021.eacl-main.115/.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018. URL https://aclanthology.org/P18-1072.

Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 4444–4451, 2017. URL http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972.

Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. Is cosine-similarity of embeddings really about similarity?, 2024. URL https://arxiv.org/abs/2403.05440.

Chihiro Taguchi, Seng Mai, Keita Kurabe, Yusuke Sakai, Georgina Agyei, Soudabeh Eslami, and David Chiang. Languages still left behind: Toward a better multilingual machine translation benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://aclanthology.org/2025.emnlp-main.1018.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. A benchmark for learning to translate a new language from one grammar book, 2023. URL https://arxiv.org/abs/2309.16575.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. URL https://aclanthology.org/D19-1449.

Feng Wang, Xiang Xiang, Jian Cheng, and Alan L. Yuille. NormFace: L2 hypersphere embedding for face verification, 2017. URL https://arxiv.org/abs/1704.06369.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report, 2024. URL https://arxiv.org/abs/2402.05672.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. Expanding pretrained models to thousands more languages via lexicon-based adaptation, 2022. URL https://arxiv.org/abs/2203.09435. ACL 2022.

Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. URL https://aclanthology.org/D19-1077.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. LexC-Gen: Generating data for extremely low-resource languages with large language models and bilingual lexicons. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. URL https://aclanthology.org/2024.findings-emnlp.818.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1703.06114.

A Implementation Details

Main model. 36-layer Transformer encoder with M2M100 layout (Fan et al., 2021), model dimension 1024, feed-forward dimension 8192, mean pooling, and a linear head producing a mean and a softplus variance (2×1024 outputs). The concept and frequency MLPs have four hidden layers of width 4096. Concept vectors are 300-dimensional and frozen; the unknown vector and the 9,448 learned symbol embeddings are trained. The model has 0.87B trainable parameters, excluding the frozen concept table. It was trained on lowercased English NLLB sentences with AdamW, learning rate 1×10^{-4} and batch size 192 on one H100, and we use the checkpoint with the lowest held-out loss (step 371K).

Training ablations. One A100 per run, identical data order, batch size 128, AdamW with learning rate 1×10^{-4}, gradient clipping at 1.0, no learning-rate schedule, no learned symbol table. MSE runs use bfloat16 autocast and NLL runs float32. Matched-step xsim for NLL against MSE: 45.1/5.9 at 10K steps, 51.7/31.6 at 20K, 54.8/47.4 at 50K, 56.0/54.6 at 100K, 56.9/56.1 at 150K and 55.8/56.9 at 200K. Without positional embeddings: 32.4/0.2, 47.1/3.0, 48.0/25.2, 47.1/41.4, 48.8/48.4 and 49.5/51.0.

Evaluation. English-anchored retrieval with cosine similarity; the xSIM++ distractor pool is the official one for each split. Decoding uses a public port of the SONAR text decoder with greedy search and chrF++ (word order 2).

B Tokenization

For 164 of the 216 devtest languages (169 of 224 dev), text is split with an existing word-level tokenizer. For the other 52 (55 on dev), almost all in non-Latin scripts, we apply a language-agnostic fallback: lowercase the text, split every punctuation mark and symbol (Unicode categories P and S) into its own token, split every Han ideograph into its own token, treat the Tibetan tsheg as a separator, and split on whitespace. Scripts that do not mark word boundaries (Thai, Lao, Khmer, Myanmar) therefore keep space-delimited phrases as tokens, which the lexicon rarely contains. The tokenization source of every language is part of the released results.

C Lexicon Construction

Propagation. Starting from the standardized Numberbatch 19.08 vectors, each iteration computes a vector for every PanLex meaning as the median of the known vectors of its member words, then predicts a vector for every PanLex word from the set of its meanings' vectors with the set encoder (12 layers, 8 heads, width 512, mean pooling), trained to reproduce held-out Numberbatch vectors. Words with a Numberbatch vector keep it. The space grows from 19.9 million vectors after the first iteration to 22.6 million after the fifth. The vectors for evaluation are then selected for the FLORES+ word list: exact matches first (364,818 after one iteration, 371,398 after five), then matches of the language-stripped word form (903,794 and 938,040), then the mappings coming from the stage of LLM retrieval from unstructured sources (22,829). A word that has no vector under its own language code receives the vector of the identical English, Spanish, French or German word form, when one exists.

Harvesting into PanLex. Reference documents (dictionaries, word lists and grammars) were collected per language by web search. A pipeline of LLM calls classifies each document, extracts headword and gloss pairs page by page, and adjudicates each candidate in a separate verification call; only verified pairs are kept. The 553,651 verified entries from 643 documents in 379 languages were linked to PanLex meanings, adding 635,153 word entries to the 79.1 million of the original graph, 280,763 new meanings, and new members for 116,934 existing ones.

Retrieval from unstructured material. For words still missing, a retrieval-grounded LLM notebook holding the language's reference documents is queried in batches of up to 75 words. The prompt requires strict grounding in the documents, omits words whose meaning is not supported, and asks for each sense a bracketed group of glosses in several high-resource languages in the form `lang/word_with_underscores`. A word receives the vector of its first gloss present in the concept space. The "without the final LLM stage" row of Table 3 drops these entries; since they are written after the exact and form-matched vectors, that lexicon equals the full one minus the retrieved entries.

D Additional Results

Decoding. The 29 languages of §5.4 were chosen by hand (not sampled) to span scripts, families and coverage levels, and decoding is greedy. The student's share of the teacher's chrF++ is 53% for the 14 languages with at least 85% coverage and 20% for the five below 50%. The gap between chrF++ and BLEU suggests that decoded sentences share words and word fragments with the reference but miss their order and fluency. For the intended use, bitext mining followed by conventional MT training, retrieval quality remains the relevant measure.

Vocabulary overlap with languages of SONAR. SONAR's language list is the set of 202 language codes of its tokenizer. For each FLORES+ language we split its devtest text into words (NFKC, lowercase, runs of letters; Han, kana, Thai, Lao, Khmer and Myanmar split per character) and measure the share of running words whose form occurs in the dev text of at least one SONAR language (union overlap) and of the single closest one (nearest overlap). Using devtest for the target and dev for the reference keeps shared sentence content out of the measure. Twelve FLORES+ codes absent from the list are SONAR languages under another code or regional variants of one (`cmn`/`zho`, `ekk`/`est`, `fil`/`tgl`, `gug`/`grn`, `sat_Olck`/`sat_Beng`, and variants of `twi`, `apc`, `nob` and `oci`); their text joins the reference under the SONAR code. Two are SONAR languages in a new script (`arb_Latn`, `min_Arab`), and 13 are outside the list. Across the 13 languages outside the list it does predict the teacher's score (ρ=0.86 for xsim, 0.81 for xSIM++, $p \leq 0.001$). The six with the least nearest-language overlap (under 20%) are `chv_Cyrl`, `kjh_Cyrl`, `mhr_Cyrl`, `myv_Cyrl`, `tyv_Cyrl` and `nqo_Nkoo`.

Table 5: Random deletion of known words at inference (20 languages, devtest, three seeds; mean over languages, %). The seed standard deviation is at most 0.5 points everywhere.

deleted	0%	10%	20%	30%	40%	50%	60%	80%
xsim	93.9	92.5	90.7	88.3	85.1	80.2	73.3	49.2
xSIM++	47.4	43.9	40.4	37.0	33.1	28.9	23.9	13.6

Deletion slopes. The within-language slope of xSIM++ on effective coverage is 0.42 (95% interval 0.38 to 0.45) for deletion rates up to 30%, 0.44 (0.41 to 0.46) up to 50%, and 0.50 (0.47 to 0.53) up to 80%. The cross-language coefficient is 0.56 (0.52 to 0.59).

Ten best and ten worst languages. By xSIM++ (coverage, xsim, xSIM++, %): best `zsm_Latn` (94.3, 100.0, 66.3), `ind_Latn` (95.1, 100.0, 65.6), `epo_Latn` (92.5, 100.0, 65.5), `ita_Latn` (99.1, 100.0, 64.7), `por_Latn` (97.9, 100.0, 63.6), `swe_Latn` (94.9, 99.9, 62.8), `fra_Latn` (99.5, 100.0, 62.3), `pap_Latn` (91.0, 100.0, 61.3), `glg_Latn` (93.8, 100.0, 60.2), `spa_Latn` (99.0, 100.0, 60.1). Worst: `dzo_Tibt` (3.3, 0.2, 0.0), `bod_Tibt` (5.3, 1.6, 0.0), `jpn_Jpan` (11.5, 0.9, 0.1), `sat_Olck` (10.4, 2.1, 0.2), `min_Arab` (13.3, 2.2, 0.2), `azb_Arab` (15.5, 2.5, 0.2),

`cmn_Hant` (9.2, 2.9, 0.2), `aeb_Arab` (15.2, 3.2, 0.2), `nqo_Nkoo` (13.1, 2.4, 0.3), `bjn_Arab` (13.1, 2.8, 0.3). All ten worst use the fallback tokenizer or a script without word spacing.

E PROBE DETAILS

Order: perturbation. Shuffling the English side of FLORES+ devtest with a fixed seed lowers the frozen teacher's mean xsim from 99.0% to 97.7% over 189 languages; the median language loses 0.07 points and 20 weak-retrieval languages lose more than five. Scored with xSIM++ against a distractor pool shuffled with the same permutation, the same perturbation costs 51.2 points (82.3% to 31.1%); all 189 languages lose more than five points and 184 more than 25. The En–Es pair stays at 100.0 xsim while its xSIM++ falls from 90.5 to 33.5.

Order: construction. Two English students are trained from scratch on SONAR's own subword tokenizer, identical except that one has no positional embeddings, which makes it invariant to word order (its self-cosine under permutation is exactly 1.000). Against the teacher's embeddings of the same English sentences, the order-blind student reaches a paired cosine of 0.862 on held-out NLLB text and 0.683 on FLORES+, against 0.958 and 0.947 with positions; both sit at the xsim ceiling, while xSIM++ of each student's English embeddings against the teacher's English embeddings on FLORES+ separates them (88.3 against 98.4). The gap grows with sentence length: the order-blind student falls from 0.882 at ten tokens or fewer to 0.649 above sixty.