ENTROPY2VEC: Crosslingual Language Modeling Entropy as End-to-End Learnable Language Representations

Patrick Amadeus Irawan[1*], **Ryandito Diandaru**[1*]
Belati Jagad Bintang Syuhada[2*], **Randy Zakya Suchrady**[3*]
Alham Fikri Aji[1], **Genta Indra Winata**[4], **Fajri Koto**[1], **Samuel Cahyawijaya**[5*]
[1]MBZUAI [2]Universitas Indonesia [3]NTU [4]Capital One [5]Cohere
{patrick.irawan, ryandito.diandaru}@mbzuai.ac.ae
[*]Main authors

Abstract

We introduce ENTROPY2VEC, a novel framework for deriving cross-lingual language representations by leveraging the entropy of monolingual language models. Unlike traditional typological inventories that suffer from feature sparsity and static snapshots, ENTROPY2VEC uses the inherent uncertainty in language models to capture typological relationships between languages. By training a language model on a single language, we hypothesize that the entropy of its predictions reflects its structural similarity to other languages: Low entropy indicates high similarity, while high entropy suggests greater divergence. This approach yields dense, non-sparse language embeddings that are adaptable to different timeframes and free from missing values. Empirical evaluations demonstrate that ENTROPY2VEC embeddings align with established typological categories and achieved competitive performance in downstream multilingual NLP tasks, such as those addressed by the LinguAlchemy framework.

1 Introduction

Linguistic typology provides a framework for classifying languages based on shared structural features, offering insights into language universals and diversity. Databases like the World Atlas of Language Structures (WALS) (Haspelmath, 2005), AUTOTYP (Bickel and Nichols, 2002), URIEL (Littell et al., 2017), and URIEL$^+$ (Khan et al., 2025) catalog these features, serving as valuable resources for researchers and practitioners in the field of computational linguistics and beyond. However, these inventories face significant limitations: they often cover only a subset of languages, leading to missing values, and they represent static snapshots of linguistic knowledge, neglecting the dynamic and evolutionary nature of languages.

Recent advancements in neural language modeling have enabled the extraction of continuous representations of languages through pre-trained models.

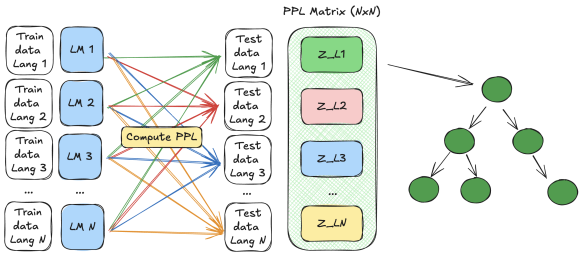


Figure 1: ENTROPY2VEC framework. Monolingual LMs are trained per language, and cross-lingual perplexity is used as an unsupervised signal to derive language vectors and induce typological trees, aligning well with expert-curated taxonomies.

These embeddings capture semantic and syntactic properties, facilitating cross-lingual transfer in various NLP tasks. Nonetheless, existing methods primarily focus on monolingual or bilingual settings and do not explicitly model the typological relationships between languages. Moreover, they often rely on manually curated features, which may not generalize well across languages or over time.

To address these challenges, we propose ENTROPY2VEC, a framework that derives language representations based on the entropy of monolingual language models (LMs). Entropy, a measure of uncertainty in information theory, reflects the predictability of a language's structure. By training a language model on a single language and analyzing its entropy when applied to other languages, we can infer typological similarities and differences. This approach offers several advantages: it is data-driven, scalable, and inherently adaptable to new languages and evolving linguistic features.

In this paper, we demonstrate that ENTROPY2VEC embeddings align with established typological categories, such as phonological, morphological, and syntactic features. We also show that these embeddings outperform traditional typological inventories in downstream multilingual NLP tasks, including language identification, typol-

Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 426–437
November 8-9, 2025

ogy prediction, and cross-lingual transfer. By integrating ENTROPY2VEC into the LinguAlchemy framework (Adilazuarda et al., 2024), we achieve competitive generalization across languages, especially those underrepresented in existing typological resources.

2 Related Works

Typological Language Inventories Traditional typological inventories, such as WALS (Haspelmath, 2005), AUTOTYP (Bickel and Nichols, 2002), URIEL (Littell et al., 2017), and URIEL$^+$ (Khan et al., 2025), have been instrumental in documenting linguistic diversity and informing computational models. However, these resources are limited by their static nature and the incomplete coverage of the world's languages. For instance, WALS provides typological data for only a fraction of the estimated 7,000 languages, leading to missing values that can hinder the performance of NLP models . ENTROPY2VEC addresses these limitations by deriving LMs from the entropy of monolingual LMs. This approach is inherently dynamic, as it can adapt to new languages and evolving linguistic features without the need for manual curation. Moreover, it provides dense, non-sparse embeddings that capture the probabilistic structure of languages, offering a more nuanced understanding of typological relationships.

Language Vector in NLP Language vectors, or embeddings, have become foundational in modern NLP, enabling models to represent words, sentences, and even entire languages as continuous vectors in a high-dimensional space. Techniques like Word2Vec, GloVe, and FastText have demonstrated that such embeddings capture semantic and syntactic properties, facilitating tasks like word similarity, analogy reasoning, and machine translation. These embeddings are typically learned from large corpora and reflect the statistical patterns in language use. However, they often treat languages as isolated entities, without explicitly modeling the relationships between them. Recent advancements, such as multilingual BERT and XLM-R, have sought to address this by training models on multiple languages simultaneously, capturing shared structures and enabling cross-lingual transfer.ENTROPY2VEC contributes to this landscape by offering a novel perspective on language representation. Instead of relying solely on large-scale pre-training on vast corpora, ENTROPY2VEC leverages the entropy of monolingual LMs to infer typological relationships between languages. This approach not only aligns with existing language representation models but also extends their capabilities by incorporating typological insights, thereby enhancing multilingual understanding and transfer learning

3 ENTROPY2VEC

3.1 Unsupervised Language Modeling

Unsupervised language modeling uses an autoregressive approach, where the LM predicts the next token based on the previous ones. Mathematically, given a sequence of tokens $[x_1, x_2, \ldots, x_t]$, the LM defines a probability distribution over the next token x_{t+1} conditioned on all previous tokens. This can be formally expressed as:

$$x_{t+1} = \arg\max_x P(x \mid x_1, x_2, \ldots, x_t; \theta)$$

where and θ represents the parameters of the model. The goal of training is to maximize the likelihood of the observed data, which is equivalent to minimizing the cross-entropy loss. Formally, given a dataset $\mathcal{D} = (x_1^{(1)}, \ldots, x_{n_1}^{(1)}), \ldots, (x_1^{(N)}, \ldots, x_{n_N}^{(N)})$, the cross-entropy loss is defined as:

$$\mathcal{L}(\theta, \mathcal{D}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{n_i} \log P(x_t^{(i)} \mid x_1^{(i)}, \ldots, x_{t-1}^{(i)}; \theta)$$

This encourages the model θ to assign high probability to the actual next tokens in the training data. The autoregressive nature of these models allows them to generate coherent and contextually relevant text by sequentially predicting tokens (Radford et al., 2019; Brown et al., 2020; Cahyawijaya et al., 2021), making them highly effective for building strong language representations (Workshop et al., 2023; Cohere et al., 2025).

3.2 LM Entropy as Language Vectors

Although having a strong language representation, LMs can only produce meaningful representation on languages that they have been pre-trained on (Winata et al., 2023; Cahyawijaya et al., 2023c) ans closely similar languages (Cahyawijaya et al., 2023b, 2024). The cross-lingual generalization often diminish when the corresponding model is faced with languages that are low-resource (Bang et al., 2023; Cahyawijaya et al., 2023a) and distant from the languages it has been trained on (Lovenia et al., 2024; Cahyawijaya, 2024; Bean et al., 2024).

As the cross-lingual generalization of LMs depends on the closeness of the language, we argue that this limitation can actually be exploited to build a language vector which is a vector that provides a global representation of a certain language. More specifically, using a set of monolingual LMs $\{\theta_{L_1}, \theta_{L_2}, \ldots, \theta_{L_n}\}$ each trained on a specific language L_i and a set of monolingual corpora $\{\mathcal{D}^{L_1}, \mathcal{D}^{L_2}, \ldots, \mathcal{D}^{L_n}\}$, we build the vector representation of languages $\{Z^{L_1}, Z^{L_2}, \ldots, Z^{L_n}\}, Z^{L_i} \in \mathbb{R}^n$ by computing the average cross-entropy of the corresponding language model θ_i on each corpus $\mathcal{D}_j$. Formally, we define the language vector Z^{L_i} as:

$$Z^{L_i} = [\mathcal{L}(\theta_i, \mathcal{D}_1), \mathcal{L}(\theta_i, \mathcal{D}_2), \ldots, \mathcal{L}(\theta_i, \mathcal{D}_n)]$$

We call our method of deriving language vector from the entropy of LMs as ENTROPY2VEC. Unlike other existing language vectors like URIEL (Littell et al., 2017) and URIEL$^+$ (Khan et al., 2025), which derive their language vectors from various linguistic inventories, e.g., WALS (Dryer and Haspelmath, 2013), AUTOTYP (Bickel et al., 2023), etc., our method provides a fully unsupervised, data-driven approach for building a language vector. Moreover, our vector can evolve following the actual evolution of languages by updating each of the monolingual LMs with more recent data on each of the corresponding languages. ENTROPY2VEC leverages the inherent patterns and structures within large-scale textual data, eliminating the need for manual feature engineering or reliance on predefined linguistic inventories. By continuously updating the models with new data, our approach ensures that the language vectors remain dynamic and reflective of the ever-changing nature of human language.

4 ENTROPY2VEC and Language Typology

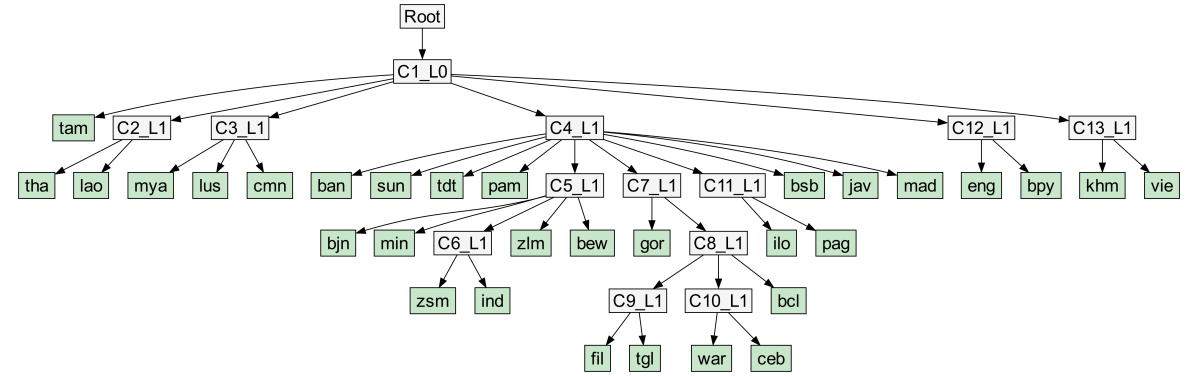
To assess the validity of ENTROPY2VEC, we compare it against several established language vector and tree baselines: URIEL (Littell et al., 2017) and URIEL$^+$ (Khan et al., 2025) vectors, as well as the Glottolog tree (Nordhoff and Hammarström, 2011). For the first two, we derive a hierarchical clustering tree representing inter-language distances based on geographical and syntactic features. We then evaluate how well the trees induced from ENTROPY2VEC vectors replicate these known typological groupings, and whether they reveal novel or diverging relationships.

4.1 Experiment Setting

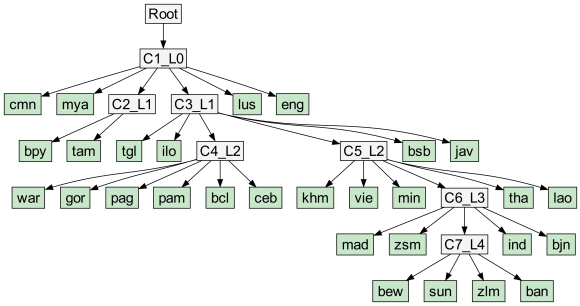
Dataset Our data source is the Glot500c corpus (Imani et al., 2023), from which we gather textual data for 33 distinct languages which are also present in URIEL, URIEL$^+$, and Glottolog. For each language, we cap the data at a maximum of 1M sentences and split this data into 7:2:1 (train, validation, test) split after collating the sentence to cap each instance to 1024 characters to support model's max ingestion length. The details of the quantity and split per language can be observed in Appendix A.

Training Strategy We choose GPT-2 as our pretrained language model for learning language representations, where the model is configured with an embedding dimension of 512, 4 transformer layers, and 8 attention heads. More details—including tokenizer configurations, optimization parameters, and the precise methodology for perplexity extraction—are elaborated further in Appendix B. Training is conducted by using the same settings for all 33 languages to extract their perplexity, a measure of how well the language model predicts the test data. This perplexity scores, reflecting the model's "surprise" by a language's characteristics, are used to derive language vectors denoted as $\{Z^{L_1}, Z^{L_2}, \ldots, Z^{L_n}\}$, where each Z^{L_i} represents a specific language. From now on, the entirety of these vectors will be termed as ENTROPY2VEC.

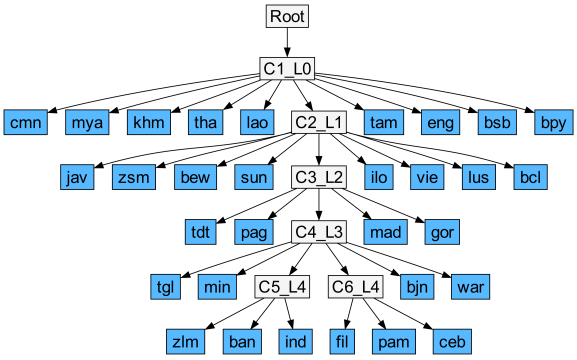
Forming Typological Trees We generate hierarchical language clusters from the learned vector representations Z^Lusing the DBSCAN algorithm, selected for its ability to discover clusters of arbitrary shape without requiring a predefined number of clusters. This choice is motivated by the non-uniform density and structure of real-world language typologies, which traditional linkage-based methods fail to capture due to its complexity (Appendix C). The resulting clusters are then transformed into tree structures and post-processed to ensure compatibility with downstream evaluation. This includes standardizing hierarchical level labels (e.g., **`family`**, **`subfamily`**, and **`language`** in URIEL and URIEL$^+$) to maintain parent-child relationships naming convention consistency. We set the same clustering hyperparameters for all experiments to **`min_samples`** = 0.3 and **`epsilon`** = 0.1. We apply these settings to all of our vector variations, including the pure ENTROPY2VEC and the its concatenated variants with URIEL or URIEL$^+$.



(a) Glottolog Tree (pruned)



(b) Tree from URIEL `syntax-knn` (pruned)



(c) Tree from ENTROPY2VEC

Figure 2: Tree comparison across methods: (a) Glottolog gold tree, (b) tree derived from URIEL `syntax-knn` distances, and (c) our tree derived using ENTROPY2VEC representations with perplexity-based clustering.

We compare the similarity of the tree generated from different language vectors with the ground truth typological tree from Glottolog [1].

Evaluating Typological Tree We extract tree subsets of the Glottolog and URIEL trees corresponding to only 33 languages present in our evaluation. Then, we use them as the gold comparison against the created ENTROPY2VEC typological tree and evaluate them using two tree distance metrics: Robinson-Foulds (RF) (Robinson and Foulds, 1981) distance and Lowest Common Ancestor (LCA) (Aho et al., 1973). The RF distance quantifies the dissimilarity between two trees based on the number of differing splits, while LCA measures the structural proximity of node pairs by comparing the depths of their lowest common ancestors. Together, these metrics assess both the global and local alignment of the induced trees.

We also report results across multiple vector concatenation ENTROPY2VEC variations (denoted as Φ), and conduct qualitative analysis to interpret the effectiveness of each representation in capturing linguistic typology.

4.2 Result and Analysis

Alignment with Language Typology Figure 2c shows the reference typological tree from Glottolog, the typological tree generated using URIEL with syntax features, and the typological tree from ENTROPY2VEC. Although there are several differences, the constructed clusters within the tree showcase the correct similarity between languages where language that come from different language families – i.e., English (eng), Tamil (tam), Chinese (cmn), and Bishnupriya Manipuri (bpy) – form their own branch on the top-level grouping, while languages that are closely similar like the Malayic language group (Hudson, 1970) – i.e., Indonesian (ind), Malay (zlm), Balinese (ban), Banjarese (bjn), and Minangkabau (min) – are grouped together. Furthermore, similar to the typological tree from

[1] Note that, there are other typological tree beside Glottolog such as Ethnologue (Campbell and Grondona, 2008) which have some differences on their typological clusters. However, as the general clusters are mostly similar, we only use Glottolog as the ground truth within our study.

Language Vector	Glottolog	
	MAE (↓)	RF (↓)
Language Features		
URIEL$_{\texttt{Geo}}$	11.11	**13.0**
URIEL$_{\texttt{Syntax}}$	9.35	18.0
URIEL$^{+}{}_{\texttt{Geo}}$	11.15	**13.0**
URIEL$^{+}{}_{\texttt{Syntax}}$	11.15	**13.0**
ENTROPY2VEC	**8.60**	17.0
Concatenated Features		
URIEL$_{\texttt{Geo}}$ + Ours	9.64	13.0
URIEL$_{\texttt{Syntax}}$ + Ours	8.58	16.0
URIEL$^{+}{}_{\texttt{Geo}}$ + Ours	**7.88**	19.0
URIEL$^{+}{}_{\texttt{Syntax}}$ + Ours	10.12	**12.0**

Table 1: Comparison of tree distance metrics between various language vector configurations and the Glottolog baseline tree. Lower MAE values and lower RF scores indicate better tree reconstruction quality. **Ours** refers to ENTROPY2VEC vectors, while "+ Ours" indicates feature concatenation with min-max normalization.

Language	ISO639-3	Family	Script	Resource
Seen Languages				
English*	eng	Indo-European	Latn	HRL
Vietnamese*	vie	Austroasiatic	Latn	HRL
Indonesian*	ind	Austronesian	Latn	HRL
Thai*	tha	Kra–Dai	Thai	HRL
Tamil*	tam	Dravidian	Taml	LRL
Burmese*	mya	Sino-Tibetan	Mymr	LRL
Ilocano	ilo	Austronesian	Latn	LRL
Javanese†	jav	Austronesian	Latn	LRL
Minangkabau	min	Austronesian	Latn	LRL
Sundanese	sun	Austronesian	Latn	LRL
Cebuano	ceb	Austronesian	Latn	LRL
Tagalog†	tgl	Austronesian	Latn	LRL
Standard Malay†	zsm	Austronesian	Latn	LRL
Unseen Languages				
German*	deu	Indo-European	Latn	HRL
French*	fra	Indo-European	Latn	HRL
Hindi*	hin	Indo-European	Deva	HRL
Italian*	ita	Indo-European	Latn	HRL
Spanish*	spa	Indo-European	Latn	HRL
Lao	lao	Kra–Dai	Laoo	LRL
Khmer*	khm	Austroasiatic	Khmr	LRL
Banjar	bjn	Austronesian	Latn	LRL
Balinese	ban	Austronesian	Latn	LRL
Mizo (Lushai)	lus	Sino-Tibetan	Latn	LRL
Waray	war	Austronesian	Latn	LRL
Buginese	bug	Austronesian	Latn	LRL
Pangasinan	pag	Austronesian	Latn	LRL
Acehnese	ace	Austronesian	Latn	LRL
Sanskrit	san	Indo-European	Deva	LRL
Fijian	fij	Austronesian	Latn	LRL
Telugu*	tel	Dravidian	Telu	LRL
Tok Pisin	tpi	Creole	Latn	LRL
Marathi	mar	Indo-European	Deva	LRL

Table 2: Detailed list of languages used in the seen and unseen evaluation in SIB-200. $*$ the language is used in MASSIVE in the corresponding subset. † the languages is used as part of unseen language evaluation in MASSIVE.

Glottog, most of the Phillipine languages (Reid and Liao, 2004) – i.e, Tagalog (tgl), Cebuano (ceb), Waray Waray (war), Pampanga (pam), and Pangasinan (pag) –, also clustered together with the Malayic due to the shared morphosyntactic features between the two groups.

We further quantify the similarity distance between these typological trees and the Glottolog ground truth typological tree as described in §4.1. The distance measures from the hierarchical clustering trees generated= from different language vectors are shown in Table 1. These metrics indicate how well the generated typological trees align with the typological tree from Glottolog. Overall, the results demonstrate that tree from ENTROPY2VEC, URIEL, and URIEL^{+} have similar alignment to Glottolog, where ENTROPY2VEC yields best LCA MAE with slightly lower RF scores in comparison to URIEL and URIEL^{+} vectors, indicating that ENTROPY2VEC captures key linguistic relationships similar to these vectors without supervision.

Combination of Language Features We also compare the base representation (Z^L) with concatenated features ($\Phi(A, B)$). Across MAE and RF, we observe that concatenation does not consistently yield improvements. Although some combinations show slight gains, others show worse performance. For example, the combined ENTROPY2VEC and URIEL^{+}Geo variant achieves the lowest MAE (7.88), indicating a closer approximation to the reference tree in terms of distances between the edges. Conversely, the combined ENTROPY2VEC and URIEL^{+}Syntax variant produces the best RF score (12.0), reflecting fewer topological errors. However, these improvements are not synergic across both metrics, suggesting that combining features may introduce redundancy or conflicting signals rather than complementarity.

Dissimilarity to Language Typology Despite the similarity, there are still some inconsistencies between trees and measurement of the distance between the expected ground truth typological tree from Glottolog and comparing the similarities and differences between different typological trees generated from different language features are not straightforward. While our ENTROPY2VEC tree broadly reflects syntactic and geographical relationships, several misalignments persist, as shown in

Language Vectors	OVR Avg.	SIB-200					MASSIVE				
		Seen		Unseen			Seen		Unseen		
		HRL	LRL	HRL	LRL	Avg.	HRL	LRL	HRL	LRL	Avg.
XLM-R											
$\text{URIEL}_{\text{Geo}}$	77.71	79.3	78.3	79.8	77.7	78.8	80.48	76.11	75.09	74.94	**76.7**
$\text{URIEL}_{\text{Syntax}}$	77.19	78.9	77.9	78.5	77.2	78.1	80.19	75.56	74.68	74.48	76.2
$\text{URIEL}^{+}_{\text{Geo}}$	77.56	79.9	78.6	80.7	77.9	79.3	79.89	75.15	74.25	74.03	75.8
$\text{URIEL}^{+}_{\text{Syntax}}$	**79.07**	82.4	81.3	82.8	80.7	**81.8**	80.16	75.71	74.85	74.70	76.4
ENTROPY2VEC (Ours)	<u>79.06</u>	82.3	81.0	82.6	80.4	<u>81.6</u>	80.31	76.16	75.00	74.73	<u>76.6</u>
$\text{URIEL}_{\text{Geo}}$ + Ours	76.72	78.2	77.2	77.6	76.4	77.3	80.12	75.36	74.46	74.42	76.1
$\text{URIEL}_{\text{Syntax}}$ + Ours	78.85	82.1	80.7	82.3	79.9	81.3	80.50	75.84	74.86	74.67	76.5
$\text{URIEL}^{+}_{\text{Geo}}$ + Ours	77.47	80.5	79.0	81.3	78.1	79.7	79.34	74.69	73.44	73.32	75.2
$\text{URIEL}^{+}_{\text{Syntax}}$ + Ours	78.78	81.7	80.7	82.1	80.2	81.2	80.30	75.84	74.80	74.57	76.4

Table 3: Accuracy comparison of different language vectors for LinguAlchemy regularization on the XLM-R backbone, using SIB and MASSIVE benchmark averages. **Bold** numbers indicate the best average performance, while <u>underlined</u> numbers indicate the second-best. We report overall performance across different settings, including seen and unseen languages during training, as well as **H**igh- vs. **L**ow-resource languages. For XLM-R, we observe that vector concatenation does not increase performance compared to their standalone counterparts, as discussed detail in subsection 5.2

Figure 2c. For example, in the predicted tree, **lao** is grouped with **tam** and **tha** under <u>Unsplit_L3_1</u> cluster node rather than with its expected Mainland Southeast Asian cluster (**vie**, **khm**) as appears in the gold-standard <u>Unsplit_L1_2</u>. Regarding the cluster sensitivity, **bpy** appears in a broad mixed group (<u>Cluster_L1_5</u>) with **jav**, **bsb**, and **war**, rather than with Tibeto-Burman-influenced languages like **lus** and **mya** as in the gold-standard <u>Unsplit_L1_3</u>. Similarly, the Malayic languages **min**, **zlm**, and **zsm** are dispersed across different branches instead of being tightly grouped under a single parent, as in <u>Unsplit_L1_1</u>. These suggest that there still lies a challenge in maintaining the persistence yntactical or geographical relationships between language groups at more granular level.

5 ENTROPY2VEC as Language Vectors

In the previous section, we demonstrate that ENTROPY2VEC is able to represent meaningful linguistic properties such as language family relation, syntax similarity, and geographical distance. In this section, we establish the applicability of ENTROPY2VEC and compare it to other existing language vectors such as URIEL (Littell et al., 2017) and URIEL^{+} (Khan et al., 2025). We compare the effectiveness of ENTROPY2VEC and other language vectors by measuring the LMs performance when applying the vectors on downstream tasks.

5.1 Experiment Setting

Training Strategy To evaluate the downstream effectiveness of ENTROPY2VEC, we utilize ENTROPY2VEC as a language vector to regularize the LMs during the fine-tuning process with LinguAlchemy (Adilazuarda et al., 2024). LinguAlchemy utilize language vectors to bring better cross-lingual generalization for low-resource and unseen languages. In this case, the downstream improvement on the low-resource and unseen languages with LinguAlchemy can be attributed to the quality of the language vector.

Dataset We incorporate SIB-200 (Adelani et al., 2024) and MASSIVE (FitzGerald et al., 2023) as our evaluation dataset. In our evaluation, we filter out the training and evaluation data to only cover the languages that are related to our 33 supported languages. This yields 13 languages for training and seen-language evaluation with additional of 19 languages for unseen evaluations for SIB-200; and 6 languages for training and seen-language evaluation with additional of 10 languages for unseen evaluations for MASSIVE. The list of languages covered for training and unseen evaluations are shown in Table 2.

5.2 Result and Analysis

Performance Across Different Settings This section discusses the impact of different language vectors to the quality of LMs across different language resource levels. The XLM-R results in Table

		SIB-200					MASSIVE				
		Seen		Unseen			Seen		Unseen		
Language Vectors	**OVR Avg.**	**HRL**	**LRL**	**HRL**	**LRL**	**Avg.**	**HRL**	**LRL**	**HRL**	**LRL**	**Avg.**
mBERT											
URIEL$_{\texttt{Geo}}$	67.61	69.4	70.9	72.7	70.2	70.8	72.40	65.24	60.25	59.77	64.9
URIEL$_{\texttt{Syntax}}$	67.49	68.8	70.2	72.5	69.6	70.3	72.63	65.53	60.56	60.14	64.7
URIEL$^{+}_{\texttt{Geo}}$	66.67	68.3	69.6	72.2	68.7	69.7	71.76	64.32	59.36	59.06	63.6
URIEL$^{+}_{\texttt{Syntax}}$	67.51	69.1	70.6	72.0	69.8	70.4	72.63	65.38	60.55	60.12	64.7
ENTROPY2VEC (Ours)	67.59	68.9	70.2	72.1	69.4	70.2	72.98	65.85	60.98	60.37	65.1
URIEL$_{\texttt{Geo}}$ + Ours	68.16	70.2	71.6	73.1	70.9	**71.5**	72.80	65.73	60.74	60.14	65.3
URIEL$_{\texttt{Syntax}}$ + Ours	68.29	70.2	71.5	73.2	70.6	71.4	72.92	66.09	61.09	60.59	65.7
URIEL$^{+}_{\texttt{Geo}}$ + Ours	67.87	69.9	71.0	72.9	70.1	71.0	72.72	65.57	60.64	60.12	65.3
URIEL$^{+}_{\texttt{Syntax}}$ + Ours	**68.59**	70.1	71.1	73.2	70.2	71.2	73.71	67.01	62.05	61.36	**66.5**

Table 4: Accuracy comparison of different language vectors for LinguAlchemy regularization on the mBERT backbone, using SIB and MASSIVE benchmark averages. **Bold** numbers indicate the best average performance, while underlined numbers indicate the second-best. We report overall performance across different settings, including seen and unseen languages during training, as well as **H**igh- vs. **L**ow-resource languages. For mBERT, we observe that vector concatenation is able to boost performance compared to standalone counterparts, as discussed detail in subsection 5.2,

3 indicate that ENTROPY2VEC provides competitive accuracy (81.3) compared to URIEL^{+} (81.5, the best baseline). The improvement is even more pronounced when compared to URIEL's Geo feature (78.5) and Syntax feature (78.1). The performance difference between HRL and LRL follows the trend observed in the baselines, both in seen and unseen languages.

Although the trend similarity between URIEL, URIEL^{+} and ENTROPY2VEC used with mBERT still persists, ENTROPY2VEC does not show any significant improvement (only resonating around 67. accuracy) compared to all baselines, as shown in Table 4. Furthermore, there is lack of difference in accuracy between HRL and LRL. This can be attributed to the limited representational understanding capability of mBERT compared to XLM-R, which results in minimal distinctions between different standalone language vectors (ENTROPY2VEC and baselines) and between languages with varying resource levels. Overall, our results highlight that ENTROPY2VEC represents a competitive or even superior vector regularizer compared to baseline performance.

Significance of Combining Vectors We also explore the potential of combining ENTROPY2VEC with baseline vectors to examine whether this leads to any amplifying effect. By concatenating ENTROPY2VEC with baseline vectors (e.g. URIEL$_{\texttt{Geo}}$ or URIEL$_{\texttt{Syntax}}$), we hypothesize that the combined vector may enrich the representation space: ENTROPY2VEC contributes information about language perplexity patterns, while the baseline vectors provide structural or typological cue.

In XLM-R, the combination does not provide additional benefit. For example, concatenating **Ours + URIEL$_{\texttt{Geo}}$** reduces the average accuracy to 77.2, which is below the standalone ENTROPY2VEC (81.3) and **URIEL$_{\texttt{Geo}}$** (78.5). A similar result is observed with the **Ours + URIEL$_{\texttt{Syntax}}$** concatenated vectors, yielding 80.9, which is less than ENTROPY2VEC (81.3) and **URIEL$_{\texttt{Syntax}}$** (81.5). Concatenations with URIEL^{+} variants also show similar trends. These results suggest that in XLM-R, combining vectors may introduce redundancy or even conflicting signals rather than complementary or synergistic gains, analogous to an overfit scenario.

In contrast, concatenation improves the performance in mBERT. The combination with **URIEL$_{\texttt{Geo}}$** increases the average accuracy to 68.16 compared to the standalone counterparts (67.61 for **URIEL$_{\texttt{Geo}}$** only and 67.59 for ENTROPY2VEC only). This trend is also observed in other combinations with URIEL^{+} baselines across all language features, as shown in Table 4. Our findings indicate that mBERT benefits from vector concatenation because the combined vectors provide stronger representations to compensate for the weaker language understanding of mBERT, as discussed in Subsection 5.2. Thus, ENTROPY2VEC can also be used to improve language representation by leveraging a weak multilingual model to improve performance.

Dataset	#Langs	Sparsity	Missing Features in Data	Last Update	Dynamic Inventory
WALS	260	Sparse	✓	2003	✗
AUTOTYP	1004	Sparse	✓	2013	✗
SSWL	178	Sparse	✓	2015	✗
PHOIBLE	2186	Sparse	✓	2019	✗
BDPROTO	257	Sparse	✓	2020	✗
Grambank	2467	Moderate	✓	2023	✗
APiCS	76	Dense	✓	2013	✗
eWAVE	77	Dense	✓	2020	✗
ENTROPY2VEC	33†	Dense	✗	2025	✓

Table 5: Comparison between linguistic inventories in WALS, AUTOTYP, URIEL, and URIEL+ and ENTROPY2VEC. † ENTROPY2VEC can be extended to 1000+ languages with open-access corpora (See §6).

6 Discussion

As highlighted in Table 5, a significant limitation of WALS, AUTOTYP, and other linguistic databases is their inherently static nature of inventories. They are the result of manual curation by linguistic experts, which process is both time-consuming and resource-intensive. As a result, they represent a fixed snapshot of the linguistic knowledge at that point in time and suffer from incomplete coverage of the world's languages. This static representation doesn't take into account that languages are dynamic and constantly evolving through gradual shift in syntax and the influence of language contact (Christiansen and Kirby, 2003; Fairclough, 2009; Corballis, 2017; Grenoble, 2021; Brochhagen et al., 2023). ENTROPY2VEC directly addresses this problem by providing a fully unsupervised data-driven framework. Since its language vectors are derived from the entropy of language models, they can change along with the language they represent. If a language community develops a new slang or undergoes a grammatical shift, those changes will be reflected in the new text corpora. This update can be performed using continual learning, where models are incrementally refined with new data rather than being fully retrain from scratch. ENTROPY2VEC alleviates the time-consuming process associated with manual database updates and allows for the rapid inclusions of newly documented or low-resource languages. It is also worth noting that, the current ENTROPY2VEC is only a prototype covering 33 languages. This however can be easily extended to thousands of languages, by incorporating large-scale corpora such as CommonCrawl [2], mC4 (Xue et al., 2021), Glot-500 (Imani et al., 2023), FineWeb 2 (Penedo et al., 2025), etc.

7 Conclusion

ENTROPY2VEC represents a significant advancement in the field of NLP, offering a novel, minimal human-derived knowledge and intervention approach to language representation that captures linguistic characteristics and achieves competitive cross-lingual generalization compared to baselines. By leveraging existing language models, ENTROPY2VEC is able to derive features with dynamic inventory without having to restart manual baseline-like typology studies and is free from the missing values that plague traditional typological language inventories. This adaptability and completeness make ENTROPY2VEC a powerful tool for representing languages, as demonstrated by its ability to mirror patterns observed in linguistic studies and enhance downstream NLP applications. The effectiveness of ENTROPY2VEC in improving cross-lingual generalization—both as its sole vector and when integrated with baselines—highlights its dynamic nature and compatibility with other representations. ENTROPY2VEC holds strong promise for advancing linguistic inclusion and supporting language documentation and preservation efforts, making it a valuable contribution to the field with significant implications for future research in language representation learning.

[2] https://commoncrawl.org/

Limitations

While ENTROPY2VEC offers several advantages, it is not without limitations. The quality of the embeddings depends on the availability and quality of monolingual corpora for each language. For languages with limited textual resources, the resulting embeddings may be less accurate or informative. Additionally, the entropy-based approach may not capture linguistic aspects, particularly those that are less predictable or more variable.

Secondly, Figure 2c shows that similar languages, such as **`thai`** and **`lao`**, are separated at an early stage of hiearchical cluster splitting, despite their expected common language ancestry relationship. This suggests that the representation is influenced by the encoding, causing similar languages to split due to differing encodings. This may not be ideal in a certain use case, as despite having different scripts, languages like **Thai**, **Khmer**, **Lao**, **Burmese** shared many vocabularies due to a closely similar geopolitical and sociocultural background (Bradley, 2009; Siebenhütter, 2019; Bradley, 2023).

Future work could integrate additional linguistic features or shared encoding structures to better capture underlying etymological relationships. Despite these challenges, ENTROPY2VEC holds promise for promoting linguistic inclusion and supporting language documentation and preservation efforts, making it a valuable contribution to the field with significant implications for future research and applications in NLP.

Ethical Consideration

The development of ENTROPY2VEC has significant implications for the field of computational linguistics and NLP. By providing a more comprehensive and adaptable representation of linguistic diversity, ENTROPY2VEC can contribute to the development of more inclusive and equitable NLP models. This can help address issues related to underrepresentation and bias in existing models, promoting fairness and accessibility in NLP applications.

However, it is essential to consider the ethical implications of using entropy-based measures to infer typological relationships. While entropy provides a quantitative measure of uncertainty, it may not fully capture the complexity and nuance of linguistic diversity. Therefore, it is crucial to complement entropy-based approaches with qualitative analyses and to remain mindful of the limitations and potential biases inherent in the data and models.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Muhammad Farid Adilazuarda, Samuel Cahyawijaya, Alham Fikri Aji, Genta Indra Winata, and Ayu Purwarianti. 2024. Lingualchemy: Fusing typological and geographical elements for unseen language generalization. *Preprint*, arXiv:2401.06034.

A. V. Aho, J. E. Hopcroft, and J. D. Ullman. 1973. On finding lowest common ancestors in trees. In *Proceedings of the Fifth Annual ACM Symposium on Theory of Computing*, STOC '73, page 253–265, New York, NY, USA. Association for Computing Machinery.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Andrew Michael Bean, Simeon Hellsten, Harry Mayne, Jabez Magomere, Ethan A Chi, Ryan Andrew Chi, Scott A. Hale, and Hannah Rose Kirk. 2024. LINGOLY: A benchmark of olympiad-level linguistic reasoning puzzles in low resource and extinct languages. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Balthasar Bickel and Johanna Nichols. 2002. Autotypologizing databases and their use in fieldwork. In *Proceedings of the international LREC workshop on resources and tools in field linguistics, Las Palmas*, volume 2627. ISLE and DOBES Nijmegen.

Balthasar Bickel, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga, and John B Lowe. 2023. The autotyp database (v1.1.1).

David Bradley. 2009. Burma, thailand, cambodia, laos and vietnam. *The Routledge handbook of sociolinguistics around the world*, pages 98–107.

David Bradley. 2023. Sociolinguistics in mainland southeast asia. In *The Routledge Handbook of Sociolinguistics Around the World*, pages 227–237. Routledge.

Thomas Brochhagen, Gemma Boleda, Eleonora Gualdoni, and Yang Xu. 2023. From language development to language evolution: A unified view of human lexical creativity. *Science*, 381(6656):431–436.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Samuel Cahyawijaya. 2024. Llm for everyone: Representing the underrepresented in large language models. *Preprint*, arXiv:2409.13897.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, and 29 others. 2023a. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, Hanung Linuwih, Bryan Wilie, Galih Muridan, Genta Winata, David Moeljadi, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2023b. NusaWrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–945, Nusa Dua, Bali. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Putri, Wawan Cenggoro, Jhonson Lee, Salsabil Akbar, Emmanuel Dave, Nuurshadieq Nuurshadieq, Muhammad Mahendra, Rr Putri, Bryan Wilie, Genta Winata, Alham Aji, Ayu Purwarianti, and Pascale Fung. 2024. Cendol: Open instruction-tuned generative large language models for Indonesian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14899–14914, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023c. InstructAlign: High-and-low resource language alignment via continual crosslingual instruction tuning. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 55–78, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lyle Campbell and Verónica Grondona. 2008. Ethnologue: Languages of the world. *Language*, 84(3):636–641.

Morten H Christiansen and Simon Kirby. 2003. Language evolution: Consensus and controversies. *Trends in cognitive sciences*, 7(7):300–307.

Team Cohere, :, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, and 211 others. 2025. Command a: An enterprise-ready large language model. *Preprint*, arXiv:2504.00698.

Michael C Corballis. 2017. Language evolution: a changing perspective. *Trends in cognitive sciences*, 21(4):229–236.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo.

Norman Fairclough. 2009. Language and globalization.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Lenore A Grenoble. 2021. Language shift. In *Oxford research encyclopedia of linguistics*.

Martin Haspelmath. 2005. *The world atlas of language structures*. Oxford University Press.

Alfred B Hudson. 1970. A note on selako: Malayic dayak and land dayak languages in western borneo. *Sarawak Museum Journal*, 18(36-37):301–318.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

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.

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.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James Validad Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius Hadiwijaya, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, and 42 others. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Sebastian Nordhoff and Harald Hammarström. 2011. Glottolog/langdoc: Defining dialects, languages, and language families as collections of resources. In *First International Workshop on Linked Science 2011-In conjunction with the International Semantic Web Conference (ISWC 2011)*.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Lawrence A Reid and Hsiu-chuan Liao. 2004. A brief syntactic typology of philippine languages.

D.F. Robinson and L.R. Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131–147.

Stefanie Siebenhütter. 2019. Sociocultural influences on linguistic geography: religion and language in southeast asia. In *Handbook of the changing world language map*, pages 2825–2843. Springer.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, and 375 others. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Dataset Split

ISO 639-3	Total Sentences	Train	Val	Test
ace	29,495	20,614	5,935	2,946
asm	1,446,686	1,012,860	289,415	144,411
ban	48,960	34,271	9,793	4,896
bcl	82,370	57,721	16,444	8,205
bew	226,176	158,323	45,235	22,618
bjn	47,158	32,997	9,425	4,736
bpy	164,807	115,282	32,999	16,526
bsb	61,759	43,228	12,350	6,181
ceb	1,433,543	1,003,516	286,718	143,309
cmn	57,500	40,250	11,500	5,750
deu	1,431,072	1,001,726	286,195	143,151
eng	1,431,047	1,001,710	286,203	143,134
fil	1,452,085	1,016,632	290,292	145,161
fra	1,430,341	1,001,232	286,082	143,027
gor	24,962	17,487	4,984	2,491
ilo	148,377	103,846	29,680	14,851
ind	1,430,227	1,001,157	286,058	143,012
ita	1,431,076	1,001,706	286,201	143,089
jav	449,862	314,774	90,134	44,954
khm	571,343	399,868	114,315	57,160
lao	56,924	39,838	11,395	5,691
lus	114,461	80,136	22,880	11,445
mad	9,055	6,055	1,500	1,500
min	593,618	415,559	118,724	59,335
mya	997,193	697,982	199,403	99,808
pag	11,812	8,268	2,365	1,179
pam	308,828	216,328	61,655	30,845
por	1,430,401	1,001,290	286,086	143,025
spa	1,430,138	1,001,097	286,027	143,014
sun	1,452,873	1,016,965	290,539	145,369
tam	1,465,996	1,026,120	293,394	146,482
tdt	7,028	4,028	1,500	1,500
tgl	1,430,721	1,001,500	286,145	143,076
tha	1,462,635	1,023,707	292,544	146,384
vie	1,436,327	1,005,431	287,358	143,538
war	1,430,401	1,001,302	286,056	143,043
zlm	30,475	21,332	6,095	3,048
zsm	849,043	594,323	169,806	84,904

Table 6: Language-wise sentence statistics with dataset splits (Train / Validation / Test). We maintain a ratio of 7:2:1 for the split, with minimum amount of 1,500 for val and test split.

B ENTROPY2VEC Training Detail

Tokenization We employ a custom character-level tokenizer. This tokenizer can either be loaded if previously trained for an experiment or trained anew on the specific language's dataset. It supports a `byte_fallback` mechanism, which, if enabled, represents characters not in the vocabulary as a sequence of their UTF-8 byte codes (e.g., "0xef"); otherwise, out-of-vocabulary characters are mapped to a `[UNK]` token. A `[PAD]` token is also utilized. During data preparation, texts are tokenized with `truncation` enabled, a `max length` of `1024` tokens, and `padding` applied to the maximum length.

More on Training Validation Evaluation is performed every `100` steps, model checkpoints are saved every `1000` steps, and a maximum of `2` checkpoints are kept. The best model, determined by the lowest eval loss, is loaded at the end of training. Both training and evaluation utilize a per-device batch size of `8`, and models are trained for up to `150` epochs. Metrics are logged every `100` steps. An `EarlyStoppingCallback` with a patience of `3` evaluations is used to prevent overfitting, and a custom `PerplexityLoggingCallback` logs perplexity during training. Data is collated for causal language modeling (i.e., mlm=False).

C Failure of Linkage-based Clustering

Traditional linkage-based clustering methods, such as agglomerative clustering with different linkage criteria (ward, complete, average) build trees by iteratively merging or splitting clusters based on simple distance metric. While effective with data with a clear, sphere-like structure, these methods fail in the context of generating language clusters due to several foundational assumptions that do not hold true for this data, which are:

Predefined Number of Clusters To derive a flat set of clusters from a linkage-based hierarchy, the number of clusters k must be specified to *cut* the dendogram. This requires the priori knowledge of the data's structure, which is often unavailable when exploring typological relationships. This methodological requirement can force an unnatural structure onto the data, potentially leading to linguistically invalid groupings.

Sensitivity to Noise and Density Variation The performance of linkage-based methods can be significantly degraded by the presence of noise and outliers. For example, single-linkage is susceptible to a "chaining" effect, where it incorrectly merges distinct clusters if a series of intermediate noise points connects them. Complete-linkage, conversely, is sensitive to outliers and may fail to merge clusters that are otherwise close.