SINFOS: A Parallel Dataset for Translating Sinhala Figures of Speech

Johan Sofalas[a], **Dilushri Pavithra**[a], **Nevidu Jayatilleke**[b] **and Ruvan Weerasinghe**[a]
[a]Research Department, Informatics Institute of Technology, Sri Lanka
{johan.s, pavithra.r, ruvan.w}@iit.ac.lk,
[b]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
nevidu.25@cse.mrt.ac.lk

Abstract

Figures of Speech (FoS) consist of multi-word phrases that are deeply intertwined with culture. While *Neural Machine Translation* (NMT) performs relatively well with the figurative expressions of high-resource languages, it often faces challenges when dealing with low-resource languages like Sinhala due to limited available data. To address this limitation, we introduce SINFOS, a dataset of 2,344 Sinhala figures of speech with cultural and cross-lingual annotations. We examine this dataset to classify the cultural origins of the FoS and to identify their cross-lingual equivalents. Additionally, we have developed a binary classifier to differentiate between two types of FoS in the dataset, achieving an accuracy rate of approximately 92%. We also evaluate the performance of existing LLMs on this dataset. Our findings reveal significant shortcomings in the current capabilities of LLMs, as these models often struggle to accurately convey idiomatic meanings. By making this dataset publicly available, we offer a crucial benchmark for future research in low-resource NLP and culturally aware machine translation.

1 Introduction

Language and culture are deeply interrelated and significant mutual influence in multiple ways (Hamidi, 2023). FoS are the tools that make language expression more vivid, attractive, and effective (Regmi, 2015). They are built through a small set of meaning-construction mechanisms where speakers reuse familiar knowledge structures in new contexts (Dancygier and Sweetser, 2014). Speakers utilise various figurative forms, such as exaggeration and idioms, as they often achieve discourse goals more effectively than literal words (Roberts and Kreuz, 1994). While idioms are universal, each language features unique expressions with specific meanings, complicating the translation process and creating a sophisticated challenge (Medagama, 2021).

The Sinhala language is part of the Indo-Aryan branch of the Indo-European language family with a rich and diverse literary heritage that has evolved over several millennia. It uses a unique script that is derived from the ancient Indian Brahmi script (Jayatilleke and de Silva, 2025b). The origins of Sinhala can be traced back to between the 3rd and 2nd centuries BCE. Sinhala is the primary language of the Sinhalese people, who make up the largest ethnic group in Sri Lanka, and it is recognised as the first language (L1) for approximately 16 million individuals (De Silva, 2025; Jayatilleke and de Silva, 2025a). According to the criteria established by Ranathunga and de Silva (2022), Sinhala is classified as a lower-resourced language (Category 2).

Sinhala has a long and well-documented tradition of FoS (භාෂා අලංකාර \ bʰɑːʃɑː ʌlʌnkʌrʌ) that appears in both literary and everyday communication (Senaveratna, 2005). They emerged gradually as Sinhala speakers and writers needed brief ways to support religious, educational, and courtly objectives, communicate indirectly and memorably in everyday conversation, and enhance the aesthetic quality of their poetry (Nawaz et al., 2025; Mieder, 1997). Currently, Sinhala FoS are mainly preserved in collections such as books and dictionaries, with many manuscripts held by national institutions and temples (Mieder, 1997). In this study, we present SINFOS [1], the first Sinhala dataset of its kind with essential data to support the task of machine translation (target language: English).

2 Related Works

A substantial body of research has examined FoS, including idioms (Sporleder et al., 2010), metaphors (Dodge et al., 2015), proverbs (Bonin et al., 2017), and other forms of figurative language (Kabra et al., 2023).

[1] https://huggingface.co/datasets/SloppyCalculator/SinFoS

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 8–26

2.1 Existing FoS Corpora

Resources are predominantly English-focused, whereas a smaller subset provides broader multilingual coverage, including European Portuguese, Danish, Chinese, and multi-language compilations such as MABL and ID10M. (Kabra et al., 2023; Tedeschi et al., 2022). The datasets ranged in size from moderate idiom/proverb collections, small lexicons (hundreds to 1,000 items) (Zhou et al., 2021; Moussallem et al., 2018), to (1̃,000–10,000) (Stowe et al., 2022; Reddy et al., 2011), with a few large-scale corpora (tens of thousands of instances/pairs or even larger textual corpora) (Zheng et al., 2019; Krennmayr and Steen, 2017). Moreover, a limited number of datasets, such as Adewumi et al. (2022), have a multi-phenomenon architecture that covers a greater variety of figurative categories, whereas many datasets are single-phenomenon resources that primarily target idioms or metaphors (Sporleder et al., 2010; Dodge et al., 2015; Prochnow et al., 2024; Shaikh et al., 2024).

Shaikh et al. (2024) introduce KonIdioms[2], an annotated Konkani idiom corpus (4,332 sentences and 817 potentially idiomatic expressions) designed to support automatic idiom identification and evaluation for this low-resource language. Furthermore, the PARSEME [3] dataset release 1.3 provides multilingual annotations of *Verbal Multiword Expressions* (VMWEs) across Arabic, Bulgarian, Chinese, Croatian, Greek, Hebrew, Hindi, Irish, Latvian, Lithuanian, Maltese, Slovene, and Turkish languages, including a dedicated category for verbal idioms alongside other VMWEs types (Savary et al., 2023). In contrast, the SemEval-2022 Task 2 dataset[4] by Tayyar Madabushi et al. (2022) focuses on idiomaticity-related modelling through sentence-level evaluation in English, Portuguese, and Galician, supporting tasks such as idiom detection and representation learning. Additionally, IMIL[5]introduces an Idiom Mapping for Indian Languages resource that links idioms across Bengali, Gujarati, Hindi, Marathi, Punjabi, Tamil, and Urdu (with English mappings), enabling cross-lingual comparison and transfer for idiom processing (Agrawal et al., 2018).

It is clear that datasets related to FoS are a significant area of focus for researchers in the field, including languages like Konkani (Shaikh et al., 2024), which falls under the same language resource category (Category 2) as Sinhala (Ranathunga and de Silva, 2022). We have also discussed various existing FoS datasets for different languages in detail in Appendix A.

2.2 Classification of FoS

Many studies have classified FoS into multiple categories, each supported by explicit definitions (Banou et al., 2025). Jang et al. (2023) categorised the FLUTE [6] dataset into four categories, such as sarcasm, similes, idioms, and metaphors. Early work, such as the *SemEval-2015 Task 11* by Ghosh et al. (2015) and the discourse-oriented analysis by Musolff (2017), primarily focused on the interplay between sentiment and specific tropes, particularly irony, sarcasm, and metaphor, in social media and public discourse. Moreover, Chakrabarty et al. (2021a) redefined figurative language data as instances of *Recognising Textual Entailment* (RTE), structuring sentence pairs that comprise a premise and a hypothesis with an associated entailment label, by drawing on five pre-existing datasets (Figurative-NLI[7] (Chakrabarty et al., 2020), datasets on *irony* compiled by Van Hee et al. (2018)[8] and Ghosh et al. (2020)[9], Sarcasm SIGN[10] (Peled and Reichart, 2017), a metaphor dataset[11] by Chakrabarty et al. (2021b)) annotated for simile, metaphor, and irony, thereby constructing a corpus of more than 12,500 RTE examples. Hayani (2016) has classified the figurative texts into 12 categories, such as metaphor, personification, hyperbole, simile, metonymy, synecdoche, irony, antithesis, symbolism, and paradox.

2.3 LLMs based Machine Translations

As mentioned by Pramodya (2023), NMT systems for low-resource, morphologically rich languages such as Sinhala increasingly adopts transfer learning and fine-tuning of multilingual sequence-to-sequence LLMs rather than SMT. As mentioned by Thillainathan et al. (2025), systematic pretraining on monolingual data followed by intermediate-task transfer provides better results than conventional single-stage fine-tuning of multilingual LLM-based MT systems in Sinhala-to-English translation. Despite these advancements, translating figurative lan-

[2] https://bit.ly/3Y4LGd3
[3] http://hdl.handle.net/11372/LRT-5124
[4] https://bit.ly/4s8k4Bt
[5] https://bit.ly/4p4SUsC
[6] https://huggingface.co/datasets/ColumbiaNLP/FLUTE
[7] https://github.com/tuhinjubcse/Figurative-NLI
[8] https://competitions.codalab.org/competitions/17468
[9] https://bit.ly/44D3O1q
[10] https://github.com/lotemp/SarcasmSIGN
[11] https://bit.ly/4rfWrGc

guage remains a challenging task. While retrieval-augmented prompting can improve the translation of idioms by offering helpful definitions or context (Donthi et al., 2025), comparative analyses show that, compared to human translations, outputs from LLMs often lack cultural nuance and tend to simplify creative metaphors (Sahari et al., 2024; Karakanta et al., 2025).

Based on existing studies, it is evident that Sinhala figurative language is underexplored in the field of computational linguistics. Incorporating this resource by identifying the dominant semantic and cultural domains reflected in Sinhala figurative language, along with translating these data from Sinhala to English, will be significant for future research. Therefore, the purpose of this work is to present a dataset of Sinhala figurative language, capture its cultural nuances, and provide an essential resource for the task of machine translation from Sinhala to English.

3 Data Collection and Annotation

The SINFOS dataset consists of 2,344 unique FoS and was compiled from a carefully curated selection of authoritative resources, including various Sinhala literary works and selected Wikipedia entries. This section provides a detailed overview of the processes involved in assembling, annotating, and preprocessing the data. An example of a record from the dataset that underwent these steps is illustrated in Figure 4 in Appendix D.

3.1 Data Assembly

A significant portion of the data, approximately 65%, was sourced from the prominent Sinhala books in this field. වාග්සම්ප්‍රදා \ vagsʌmprʌðʌ - Idioms (Department of Official Languages), අතීත වාක්‍ය දීපනිය \ ʌθi:θʌ va:kyʌ ði:pʌnɪjʌ - Atheetha Wakya Deepanya (Senanayaka, 1880), and the Dictionary of Proverbs of the Sinhalese (Senaveratna, 2005), while the remaining 35% was extracted from Wikipedia [12].To ensure high fidelity to the source material, the core Sinhala expression was collected as the primary data entry. This is a foundational practice validated by benchmarks like the IDIX (Sporleder et al., 2010) and the ChID (Zheng et al., 2019) corpora, which rely on the collection of specific linguistic expressions as the base unit for identification.

[12] https://bit.ly/4qdZy08

3.2 Annotation Process

To ensure the accuracy of the sources, the annotation process closely followed the resources outlined in subsection 3.1 and was carried out by native Sinhala speakers. Importantly, when primary sources lacked the expected information related to translations (although the attributes *Literal / Visual Image* and *Type of FoS* involved some human annotation as detailed in subsections 3.2.1 and 3.2.2), the annotators refrained from using personal knowledge to avoid potential subjective interpretations. Instead, they strictly drew from previously verified resources. For example, *What it really implies* was derived directly from the *Corresponding FoS in English* found in the source books, utilising standard references such as Merriam-Webster (Dictionary, 2002) and the Cambridge Dictionary (Brown et al., 2013) for validation. Similarly, missing *Literal Image* entries were translated strictly from the FoS text, while *Type of FoS* categories were assigned based solely on the logical frameworks outlined in subsection 4.1 and Appendices B, C. A final comprehensive review confirmed that all entries were grounded in these external standards, ensuring high data integrity. As a result of the procedures followed, certain records did not include some attributes, as shown in Table 1.

Attribute	Count
Sinhala (සිංහල \ sɪŋhʌlʌ)	2344
Type of FoS	2344
Literal / visual image	2344
Corresponding FoS in English	1571
What it really implies	2059
Additional Context	125

Table 1: Distribution of annotated fields in the dataset.

3.2.1 *Type of FoS*

To clarify the figurative language associated with each record, the dataset includes a "*Type of FoS*" attribute. This granularity was essential for determining the distinct processing strategies required for different figurative types, a necessity highlighted by the PIE corpus (Adewumi et al., 2022), which classifies data into specific types like metaphors and similes, and the IMPLI study (Stowe et al., 2022), which demonstrates that models process idioms and metaphors differently.

The entries are organised into five main categories, as detailed in Table 2. Most of the idioms were obtained from (Department of Official Lan-

Type of FoS	Number of Entries
Proverbs (ප්‍රස්තාපිරුළු \ prʌsθapɪrulu)	988
Idioms (වාග්සම්ප්‍රදා \ vagsʌmprʌðʌ)	1319
Adages (ආප්තෝපදේශ \ ɑ:pθɔ:pʌðe:ʒʌ)	15
Idiosyncratic (පුද්ගලික \ puðgʌlɪkʌ:ʒʌ)	11
Sayings (කියමන් \ kɪjʌmʌn)	11

Table 2: Distribution of Entries by Figure of Speech Type.

guages), while the majority of the proverbs were gathered from (Senaveratna, 2005). For certain FoS, specific types of FoS annotations were readily available, allowing us to directly categorise them within our classification strategy and document them accordingly. The remaining FoS were annotated based on the criteria outlined in subsection 4.1. The guidelines provided in Appendix C were used to distinguish between proverbs and idioms. Additionally, proverbs were categorised into three subcategories based on their intent, origin, and conclusion. These annotations were performed according to the criteria in Appendix B. Proverbs were assigned tags corresponding to the three categories mentioned earlier, while the other types of figurative speech were labelled directly, using their Sinhala names.

3.2.2 *Literal / Visual Image*

SINFOS uses a "*Literal / Visual Image*" annotation for each entry to provide a visual reference for non-native speakers by eliminating all abstract concepts, emotions, and symbolism. Documenting the literal imagery aligned with psycholinguistic research on imageability and methodologies for testing compositionality. Since the majority of these expressions are figurative, capturing the mental image was highly necessary. Furthermore, the inclusion of the implied meaning provided the ground truth required to test a model's ability to transcend surface definitions, mirroring the "real vs. false definition" methodology of the *Danish Idiom Dataset* (Sørensen et al., 2025).

Majority of the annotation was done using the above given sources as the relevant visual details were provided by them, whilst the others were annotated by translating the Sinhala FoS, word by word (e.g.,එක හඬින් \ ɛkʌ hʌndɪn as "With one voice"). The annotation process adhered to precise guidelines for aligning words, ensuring direct correspondence between the nouns and verbs in the original Sinhala text and their English descriptions. To maintain a "Semantic Ground Truth" and avoid introducing an outside context, only tangible objects and specific actions were documented. Furthermore, non-translatable "cultural objects" were preserved in their original form. For example වැඩි පද ගහන්නේ තොවිලෙ කැතවෙන්නයි \vædɪ pʌðʌ gʌhʌnne: θɒvɪle: kæθʌvɛnnʌj was annotated as "Too much tom-toming means that the tovila is going to be spoilt", retaining the word "තොවිලෙ \ θɒvɪlɛ - Tovila (devil-ceremony, exorcism)". This method helps prevent "translation loss" and ensures that the dataset's literal accuracy is preserved, avoiding misleading interpretations that could arise from forced or inaccurate translations of culturally specific items.

3.2.3 *Corresponding FoS in English*

The attribute "*Corresponding FoS in English*" refers to the equivalent English figurative expression (FoS) for its Sinhala counterpart. One of the techniques explored by translators is direct substitution, which effectively facilitates the understanding of figurative language across different languages, even without explicit meanings (Adelnia and Dastjerdi, 2011). This process further enabled the identification of cross-lingual equivalence and cultural parallels, a parallel alignment approach that was validated through the cross-linguistic mapping of proverbs in *PROMETHEUS* (Özbal et al., 2016) and the alignment protocols of *ParaDiom* (Donaj and Antloga, 2023).

The FoS obtained from Department of Official Languages included corresponding English FoS for all entries, whereas Senaveratna (2005) provided corresponding English FoS for only some entries, which were used for annotation. Additionally, the process of annotating this data also aided in determining the "*What it really implies*" aspect for certain FoS.

3.2.4 *What it really implies*

The "*What it really implies*" column was established to clearly explain Sinhala figurative phrases in English, capturing their deeper meaning. It translates each Sinhala figurative expression into a shared human experience. Given that recognition of FoS is highly context-dependent, additional context is included to assist in disambiguation and cultural grounding. This field captures terms specific to Sinhalese culture, regional variations, and the folklore or stories behind specific figures of speech, ensuring the dataset serves as a comprehensive resource for understanding the "naked truth" behind

the language. This is supported by the context-dependent annotation standards of *EPIE* (Saxena and Paul, 2020) and the cultural analysis frameworks of *PROMETHEUS* (Özbal et al., 2016).

To maintain clarity in the data and prevent lengthy explanations, the annotation process prioritised brevity over excessive detail. Only essential translations were included, omitting additional context or details that could complicate data analysis. Most implications in the expressions were derived directly from primary reference sources mentioned in the subsection 3.1. However, when a corresponding English equivalent was identified, the meaning was modified to align with the common interpretation of that English idiom. To guarantee reliable data, entries lacking a source-based explanation or an English equivalent were excluded. This mitigates the risk of inaccuracies or subjective misinterpretations. The annotations adhere to a specific format to aid in computational modelling. Behavioural advice and actions are expressed in the infinitive form. Character types or scenarios are described in formal terms. By eliminating secondary imagery and metaphorical elements, this approach clarifies the meaning for non-native speakers. It offers a clear "ground truth " for comparing the literal interpretation of a phrase with its actual significance.

3.3 Data Pre-processing

During the pre-processing stage, meticulous attention was devoted to punctuation, particularly in the context of FoS. The retention of punctuation marks in these instances is crucial, as they play a significant role in determining both prosody and syntactic structure, which are essential for achieving accurate processing. To ensure this dataset does not leak important information about figurative language, no further word-level or sentence-level filtration was conducted on any records, including those containing stereotypes, to facilitate authentic cultural analysis and the study of historical societal norms.

4 Analysis of SINFOS

The SINFOS dataset comprises 2,344 FoS, totalling 8,903 words. The literal image section includes 14,383 words, while the "What it really implies" section has 19,386 words. On average, each Sinhala FoS consists of 3.798 words. A brief overview of the dataset statistics is shown in Table 3.

Category	N	Mean	Median	Max	Total
Sinhala FoS	**2344**	**3.80**	**3**	**24**	**8903**
Literal / visual image	2344	6.14	5	38	14383
What it really implies	2059	9.41	8	56	19366
Corresponding FoS in English	1571	3.44	3	21	5401

Table 3: Summary statistics of word counts across different categories.

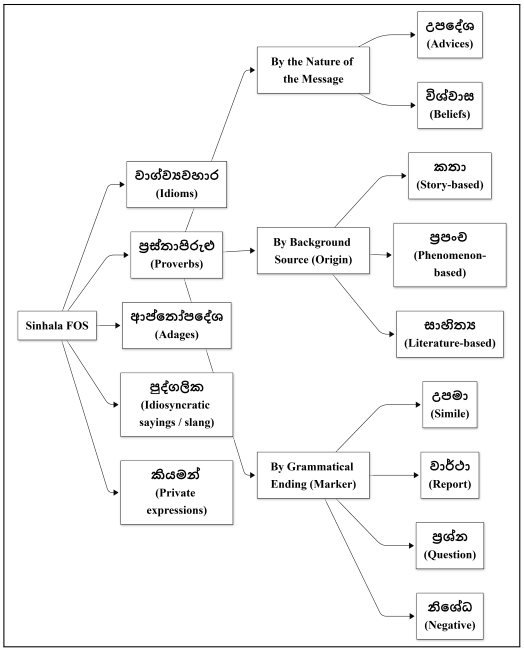


Figure 1: Summary of Sinhala FoS Dataset Classification

4.1 Classification of FoS

The classification of Sinhala FoS (භාෂා අලංකාර \ bʰɑ:ʃɑ: ʌlʌnkʌrʌ) is complex due to the fluidity of the language and its deep rooting in oral tradition. As mentioned in the subsection 3.2.1, this study classified Sinhala FoS into five main categories. The etymological roots of these terms provide a necessary framework for understanding their usage.

වාග්සම්ප්‍රදා \ vagsʌmprʌðʌ (Sinhala idioms): Derived from the Sanskrit roots " වාග්/වාක් \ vɑ:g/vɑ:k " (speech/word) and " සම්ප්‍රදාය \ sʌmprʌðɑ:jʌ " (tradition/heritage), this term refers to speech patterns established by long-standing usage. Unlike proverbs, which are often wisdom-based, these are usage-based constructs where the meaning transcends the literal definitions of the individual words. These are typically incomplete phrases or fragments, often ending in a verb. For example ආවාට ගියාට \ ɑ:vɑ:tʌ gɪjɑ:tʌ literally translates to "For coming and going" while it actually means "not friendly, and showing

little interest in other people in a way that seems slightly rude".

ප්‍රස්තාපිරුළු \ prʌsθapɪrulu (Sinhala proverbs): This is a compound of " ප්‍රස්තා \ prʌsθa: " denoting a specific occasion, moment, or opportunity, and " පිරුළු \ pɪrulu " referring to a simile, reply, or adage. Consequently, this functions as a "situational simile", a pre-packaged linguistic unit invoked to address a specific incident by comparing it to a known truth. In contrast to Sinhala idioms, Sinhala proverbs are syntactically complete sentences or clauses that can stand alone. For example ඉඟුරු දීලා මිරිස් ගන්නා වගෙයි \ ɪnguru ði:lɑ: mɪrɪs gʌθθɑ: vʌgɛj (Like exchanging ginger for chili). To provide a granular analysis, Sinhala proverbs were further classified based on the nature of the message, the source of the background, and the grammatical ending as mentioned in Appendix B.

ආප්තෝපදේශ \ ɑ:pθɔ:pʌðe:ʒʌ (Sinhala adages): Unlike figurative proverbs, these are literal directives. They represent the prescriptive aspect of the language (what one should do), distinct from the descriptive nature of idioms. An example of adages in SINFOS is ඉගෙනීම නොනැසෙන ධනයකි \ɪgʌni:mʌ nɒnæsɛnʌ ðʰʌnʌjʌkɪ (Education is an indestructible form of wealth).

පුද්ගලික \ puðgʌlɪkʌ:ʒʌ (Idiosyncratic): These are hyper-local sayings used by individuals or small groups. While not yet FoS in the public domain (Crocker, 1977), they represent the genesis point of language evolution, where personal metaphors potentially graduate into public idioms over time. Slang also falls under this category. For example the phrase අභිධර්ම මුදලාලිගෙ හෝටලේ වගේ \ ʌbʰɪðʰʌrmʌ mʊðʌlɑ:lɪge: hɔ:tʌle: vʌge: (Like Abhidharma mudalali's hotel) would be well understood by the people living in the surroundings but not by everyone.

කියමන් \ kɪjʌmʌn (Sinhala sayings): Concise verbal phrases are commonly used in daily conversation to express a thought, comment, or observation. In contrast to proverbs or idioms, these do not inherently possess a moral lesson, universal truth, or established figurative interpretation recognised by a large group. As an example මරුවා ආ දාට බාදා නැහිලු \ mʌru:vɑ: ɑ: ðɑ:tʌ ba:ðɑ: næθɪlu: (When death comes, there is no let or hindrance).

The dataset primarily consists of Sinhalese proverbs and idioms, leading to the creation of a binary classification model aimed at distinguishing between proverbs and idioms. A Voting Ensemble model, incorporating Support Vector Machines (SVM), Random Forest, and XGBoost with TF-IDF Character 3-Gram vectorisation, achieved an impressive accuracy of 90.56%. This approach, based on character-level processing, effectively tackled the intricacies of Sinhala morphology (Priyanga et al., 2017) by detecting subword elements rather than relying solely on exact phrases. The implementation of Word2Vec embeddings significantly improved performance compared to experiments based on TF-IDF (sparse vector representation). This includes the accuracy of the TF-IDF Character 3-Gram in both the Gaussian Naive Bayes and Linear SVC models, achieving an accuracy of 90.34% in each case. The analysis indicated that specific verb endings served as strong indicators of idiomatic expressions, while comparative particles and rhythmic consonant clusters were associated with proverbs. Incorporating 3-gram TF-IDF was used to leverage the identified patterns, resulting in models with these embeddings performing better than their word-level counterparts. The semantic understanding provided by dense embeddings, such as Word2Vec, also proved effective in recognising these patterns. Ultimately, utilising a Deep Feed Forward Neural Network (Deep NN), which offers superior semantic understanding, achieved the highest overall accuracy of 92.7% and the best recall for proverbs at 94%. The embeddings for the LSTM and Deep NN models are not specified in Table 8, as they relied on the standard TensorFlow Keras embeddings that learned directly from the training data.

Model	Vectorization	Accuracy	P-Rec.	I-Rec.
Gaussian Naive-Bayes	Word2Vec	90.34%	92%	89%
LinearSVC	Word2Vec	90.34%	90%	91%
Random Forets	TF-IDF (Char 3)	89.27%	83%	94%
XG-Boost	TF-IDF (Char 3)	90.13%	86%	93%
Ensemble (SVC+RFM+XGB)	TF-IDF (Char 3)	90.56%	85%	95%
Bi-LSTM	-	91.63%	86%	96%
Deep NN	-	92.7%	94%	92%

Table 4: Model Performance Comparison. Further details in Appendix F. *Note that P-Rec = Recall for Proverbs and I-Rec = Recall for Idioms.

4.2 Cultural Analysis

This research employed a hybrid methodological approach that combined both inductive and deductive thematic analysis to explore the relationship between physical imagery and cultural significance

in Sinhala FoS. This computational analysis was conducted on English translations of the dataset. The analysis identified two main aspects of the FoS: "*Literal / Visual Image*" (Source Domain), which consists of the tangible visual components that make up the figure of speech, and "*What it Really Implies*" (Target Theme), which signifies the deeper abstract or cultural meanings conveyed by the text. To minimise researcher bias and ensure that the coding frameworks were derived from raw data rather than from preconceived notions, we emphasised a bottom-up discovery phase. This inductive stage employed unsupervised machine learning methods to uncover naturally occurring patterns. Specifically, we applied TF-IDF vectorisation (using unigrams and a maximum of 2,000 features) along with K-Means clustering (k=5) to analyse the "What it Really Implies" dimension and uncover hidden linguistic clusters.

Additionally, we conducted a frequency analysis using a *Bag-of-Words* (BoW) model for both the "*Literal / Visual Image*" and "*What it Really Implies*" dimensions. This analysis allowed us to identify the most frequent and significant terms in each cluster, categorising specific words under different themes and establishing a data-driven basis for the theoretical coding frameworks. After completion of the exploratory phase, the recognised patterns were compiled into an organised dictionary for the deductive phase. We employed a rule-based classification system, using the specific keywords identified in the earlier phase as indicators of broader cultural categories. The algorithm compared the text against this predefined dictionary; if a keyword associated with a certain category was found, that category was assigned to the entry. This approach enabled multi-label classification, assuming that the subject matter remained consistent across the figurative language, thereby confirming that the detected keywords were suitable representations of the main concepts.

Lastly, a bivariate cross-tabulation was performed to quantitatively evaluate the connections and dependencies between the identified Source Domains and Target Themes. The findings reveal that Somatic (Body) and Agrarian (Nature) imagery are the most prevalent source domains, with notable mentions of the hand (n=56), water (n=46), and trees (n=43). The most frequently encountered themes are Ethics & Moral Character (n=162) and Karma & Consequence (n=127). This suggests a distinct metaphorical framework in which nature-related metaphors primarily promote moral conduct (n=20), while physical imagery specifically illustrates the tangible repercussions of karmic consequences (n=14). The distribution of literal source domains and abstract cultural themes observed in SINFOS is summarised in Table 7 in Appendix E. This implies that these FOS primarily serve as mechanisms for reinforcing social norms rather than simply providing descriptive observations.

4.3 Cross-Lingual Equivalence Analysis

This study investigates a collection of 1,571 Sinhala phrases that have English "Literal/ Visual Image" translations. This sample is derived from the initial dataset of 2,344 phrases, as the remaining 773 lack direct English equivalents. The findings indicate a notable cultural divergence, demonstrated by a symbolic overlap score of merely 0.05 using the Jaccard Index and a lexical similarity score of 0.32. The lexical similarity was calculated using the sequence matcher in the `difflib` [13] library, which employs the *Ratcliff/Obershelp Algorithm* (Ratcliff and Metzener, 1988). This implies that although the functional meanings align, the underlying metaphors originate from distinct contexts.

For example, Sinhala employs the expression "exchanging ginger for chilli," while English phrases refer to "jumping out of the frying pan into the fire." In terms of structure, 93.3% of the phrases retain their original form, while 4.9% transition from Sinhala similes into English metaphors. An illustration is "Like the eye," which transforms into "Apple of one's eye."

Furthermore, expressions in Sinhala are, on average, 32% longer than their English counterparts, yielding a ratio of 1.32. This distinction is effectively showcased by the English phrase "red herring," which in Sinhala translates to an elaborate depiction where "the fox conceals the fowl in the forest and scurries about, swinging a coconut husk from its mouth."

5 Benchmarking on LLMs

In this section, we use SINFOS as a benchmark to evaluate the performance of selected LLMs and *Small Language Models* (SLMs) in translating these complex expressions. A subset of 499 FoS was curated based on specific criteria: they represent diverse categories and possess intricate meanings that are particularly challenging for mod-

[13] https://bit.ly/4p48y7o

System/Context:

You are an expert linguist specialising in Sinhala (Sri Lanka) language and folklore.

Task:

I will provide a list of Sinhala Figures of Speech. For each item, provide only the English Figurative Meaning (what it really implies in a specific context). Do not translate literally. Do not explain the literal words.

Figure 2: Prompt used to generate responses from LMs.

els to interpret (Tayyar Madabushi et al., 2022). To ensure a comprehensive evaluation, we employed stratified sampling, purposefully oversampling rare categories, such as adages (11), "private" expressions (10), and sayings (3), which are often overshadowed by dominant idioms (190) and proverbs (285). This approach allows for a robust assessment of model capabilities across the full spectrum of figurative language, prioritising interpretative difficulty to test the distinction between literal cues and cultural nuances (Tayyar Madabushi et al., 2022). Furthermore, proverbs were broken down into their core elements (story, nature, and literature) to better analyse the depth of cultural understanding.

We used the same prompt for all models to establish a consistent evaluation baseline. Figure 2 shows the prompt provided to the *Language Models* (LMs) to elicit the meanings of the FoS. This method helps avoid prompt-induced bias, as small variations in wording could unintentionally favour one LM over another, ensuring that the responses are directly comparable.

Model	Cosine Similarity	Fidelity Scores
Gemini 3 Pro	0.6678	0.3117
Llama 4 Maverick	0.6400	0.2351
Grok 4.1	0.6354	0.2361
GPT 5.2	0.6221	0.2090
DeepSeek-V3.2	0.6126	0.2052
Claude Sonnet 4	0.5972	0.1628
Gemma	0.6024	0.1914
GPT 4.1 mini	0.5816	0.1300
Qwen 3	0.5596	0.1247

Table 5: Performance of language models on Sinhala FoS.

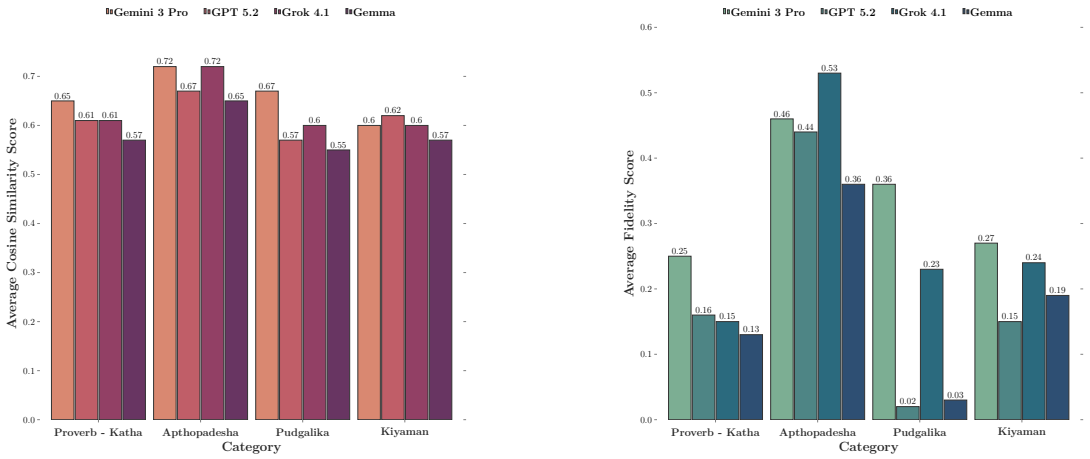
To evaluate how effectively LMs grasp FoS in Sinhala, this research employs a dual framework that examines both context retrieval and logical comprehension. This method reflects the two-step process of theme identification and truth condition mapping by Reimers and Gurevych (2019). The initial phase utilises a bi-encoder architecture with `FlagEmbedding` (specifically the *BAAI/bge-large-en-v1.5* [14] model) to calculate Cosine Similarity between the outputs of the model and the meanings annotated in the dataset. This model was selected for its state-of-the-art performance on the *Massive Text Embedding Benchmark* (MTEB), ensuring precise high-dimensional mapping that outperforms standard baselines in capturing "Semantic Relatedness" (Chen et al., 2024; Tayyar Madabushi et al., 2022).

Although this segment efficiently penalises thematic discrepancies, such as mixing "betrayal" with "love," it may be influenced by the "Keyword Bag" problem, in which comparable terms obscure gaps in logical coherence. For example, the idiom කොහොඹ ගහට කරවිල වැල ගියා වගෙයි \ kɒhɒmbʌ gʌhʌtʌ kʌrʌvilʌ vælʌ gɪjɑ: vʌge: which implies the compatible union of two negative forces (literal image: 'like the karawila creeper twining round the kohomba tree') received a high similarity score of 0.805 for the DeepSeek V3 translation, 'a mismatched or absurd pairing', despite the model's output conveying the exact opposite meaning.

To tackle this issue, the second segment measures the Fidelity Score, which implements a Cross-encoder (*stsb-roberta-large*) to evaluate intricate dependencies by analysing sentences concurrently (Reimers and Gurevych, 2019). In this context, Fidelity represents the semantic faithfulness of the model's output to the ground truth. This functions as a replacement for "Semantic Entailment," aiding in the differentiation between sentences that share similar phrasing but convey distinct meanings, such as "the dog bit the man" versus "the man bit the dog" (Li et al., 2024). By utilising the full self-attention mechanism of the Cross-encoder, the framework captures the syntactic nuances often missed by Bi-encoder models. Integrating this Fidelity Score with the first segment provides robust safeguards against "Low-Resource Hallucination," enabling a comprehensive assessment of Language Models in the Sinhala language (Benkirane et al., 2024).

At the same time, the Fidelity scores struggle with something known as the "Hyper-Literal" problem, where creative paraphrasing could be penalised. For example, the phrase බුරන බල්ලෝ හපාකන්නේ නැහැ \ bʊrʌnʌ bʌllɔ: hʌpɑ: kʌnne: næhæ is directly translated as "Barking dogs don't bite" by DeepSeek V3. In the case of translating FoS, substitution with a valid FoS

[14] https://huggingface.co/BAAI/bge-large-en-v1.5



(a) Cosine Similarity Score Comparison for Selected Categories

(b) Fidelity Score Comparison for Selected Categories

Figure 3: Benchmarking LLM Performance: (a) Cosine Similarity and (b) Fidelity Score. Information on all LLM performances could be found in Appendix G.

is considered to be a valid form of translation (Adelnia and Dastjerdi, 2011), but Fidelity gives it a modest score of 0.0089, as both phrases do not have lexical overlap. Relying only on one of these metrics can cause blind spots and skew evaluation results.

Therefore, by including both metrics, we can better assess the model's performance. This method identifies "hallucinated relevance," where high Cosine scores suggest understanding, but low Fidelity scores indicate a lack of grasp on underlying intent. This helps benchmark true understanding over mere statistical matching. Table 5 displays the average Cosine Similarity Scores and average Fidelity Scores obtained by each of these models across all the FoS available on the stratified sample obtained on the dataset based on the types of FoS, difficulty and figurativeness.

The assessment of nine advanced models reveals that `Gemini` stands out in its ability to analyse Sinhala FoS, achieving the top scores in Cosine Similarity and Fidelity. The success of the smaller `Gemma` model indicates that cultural relevance takes precedence over the model size. Nonetheless, there is an issue known as the "illusion of competence." Some models can effectively retrieve context but falter in logical comprehension. As a result, they may identify the correct domain but often misinterpret the meanings. Conventional metrics, such as BLEU, do not adequately address this challenge. Furthermore, models such as `GPT-4o mini` and `Qwen` exhibit "broken figurative triggers," offering literal interpretations instead of figurative ones for specific expressions. While most models perform well with sayings that align with Western proverbs, they tend to struggle with distinct and folklore-inspired proverbs. This stems from their literal approach to translation, which neglects the cultural context needed to understand nuances.

6 Conclusion

This study introduces SINFOS, a dataset containing 2,344 Sinhala FoS accompanied by expert-verified explanations. The annotation process is comprehensively explained in the paper. The available details were entered into the dataset, and the missing details were handled in a manner consistent with the structure of the entered details to ensure the dataset's accuracy and validity.

The analysis of the dataset emphasises a significant disparity in meaning between Sinhala idioms and their English equivalents. The cross-linguistic examination revealed the disparities among the languages, while the cultural analysis showcased the distinct culture reflected in the FoS, emphasising the challenges of translation. While LLMs can effectively handle FoS with direct English translations, they often struggle with culturally specific terminology. This can result in inaccuracies or literal conversions. Future research should focus on improving the verification of these results by implementing ablation studies and presenting statistical significance. Consequently, SINFOS serves as a vital resource for developing novel approaches in Machine Translation and modelling frameworks that seek to integrate cultural insights into languages with fewer resources.

Limitations

Sinhala meaning unavailability: A key limitation of this study is the incomplete availability of English meanings for some Sinhala FoS. In several cases, authoritative definitions or consensus interpretations were not available in accessible reference sources, which constrained some of the analysis, such as where cross-lingual analysis could not be performed across all the FoS, and the domains spoken by these FoS could not be analysed in the cultural analysis.

Meaning loss in English rendering: Some Sinhala FoS are highly culture-bound, context-dependent, or rely on implicit background knowledge, making direct English rendering difficult and increasing the risk of ambiguity or meaning loss. As a result, a portion of the dataset may contain paraphrased or approximate meanings rather than fully equivalent English interpretations, which can affect translation quality and downstream classification performance.

Class imbalance in පුද්ගලික \ puðgʌlikʌ:ʒʌ and කියමන් \ kɪjʌmʌn categories: The dataset exhibits class imbalance, particularly within the පුද්ගලික \ puðgʌlikʌ:ʒʌ and කියමන් \ kɪjʌmʌn categories, where only 11 instances were available for both categories. Therefore, the analysis done was heavily influenced by the dominant idioms and proverbs. A classification model could not be trained to classify all FoS due to the class imbalance.

References

Amineh Adelnia and Hossein Vahid Dastjerdi. 2011. Translation of idioms: A hard task for the translator. *Theory and Practice in Language Studies*, 1(7):879–883.

Tosin Adewumi, Roshanak Vadoodi, Aparajita Tripathy, Konstantina Nikolaido, Foteini Liwicki, and Marcus Liwicki. 2022. Potential idiomatic expression (PIE)-English: Corpus for classes of idioms. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 689–696, Marseille, France. European Language Resources Association.

Ruchit Agrawal, Vighnesh Chenthil Kumar, Vigneshwaran Muralidharan, and Dipti Sharma. 2018. No more beating about the bush : A step towards idiom handling for Indian language NLP. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Katsiaryna Aharodnik, Anna Feldman, and Jing Peng. 2018. Designing a Russian idiom-annotated corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Israa Alsiyat, Scott Piao, and Mansour Almansour. 2023. Arabic metaphor corpus (amc) with semantic and sentiment annotation. page 1. The twelfth International Corpus Linguistics Conference, CL2023 ; Conference date: 03-07-2023 Through 06-07-2023.

David Antunes, Jorge Baptista, and Nuno J. Mamede. 2025. A European Portuguese corpus annotated for verbal idioms. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 58–66, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Zouheir Banou, Sanaa El Filali, El Habib Benlahmar, Fatima-Zahra Alaoui, Laila El Jiani, and Hasnae Sakhi. 2025. A systematic review of figurative language detection: Methods, challenges, and multilingual perspectives. *Natural Language Processing Journal*, 13:100192.

Kenza Benkirane, Laura Gongas, Shahar Pelles, Naomi Fuchs, Joshua Darmon, Pontus Stenetorp, David Ifeoluwa Adelani, and Eduardo Sánchez. 2024. Machine translation hallucination detection for low and high resource languages using large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9647–9665, Miami, Florida, USA. Association for Computational Linguistics.

Patrick Bonin, Alain Méot, Jean-Michel Boucheix, and Aurélia Bugaiska. 2017. Psycholinguistic norms for 320 fixed expressions (idioms and proverbs) in french. *The Quarterly Journal of Experimental Psychology*, 71:1–37.

Edward Keith Brown, James Edward Miller, and James Edward Miller. 2013. *The Cambridge dictionary of linguistics*. Cambridge University Press.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021a. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020. Generating similes effortlessly like a pro: A style transfer approach for simile generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6455–6469, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021b. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4250–4261, Online. Association for Computational Linguistics.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

J Christopher Crocker. 1977. The social functions of rhetorical forms. *The social use of metaphor: Essays on the anthropology of rhetoric*, 2:33–66.

Barbara Dancygier and Eve Sweetser. 2014. *Figurative Language*. Cambridge Textbooks in Linguistics. Cambridge University Press, New York, NY, USA. Also available as paperback ISBN 978-0-521-18473-1 and PDF ISBN 978-1-107-77687-6.

Nisansa De Silva. 2025. Survey on publicly available sinhala natural language processing tools and research. *arXiv preprint arXiv:1906.02358*.

Department of Official Languages. *Idioms*. Department of Official Languages, Sri Lanka, Colombo, Sri Lanka.

Merriam-Webster Dictionary. 2002. Merriam-webster. *On-line at http://www. mw. com/home. htm*, 8(2):23.

Ellen Dodge, Jisup Hong, and Elise Stickles. 2015. MetaNet: Deep semantic automatic metaphor analysis. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 40–49, Denver, Colorado. Association for Computational Linguistics.

Gregor Donaj and Špela Antloga. 2023. ParaDiom: A parallel corpus of idiomatic texts. In *Text, Speech, and Dialogue*, page 147–158.

Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Doh, and Eid Rodan. 2025. Improving llm abilities in idiomatic translation. In *Proceedings of the 1st Workshop on Language-Oriented Research in SLMs (LoResLM)*. Also available as arXiv:2407.16470.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Eirini Florou, Konstantinos Perifanos, and Dionysis Goutsos. 2018. Neural embeddings for metaphor detection in a corpus of Greek texts. In *2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–4.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–478, Denver, Colorado. Association for Computational Linguistics.

Debanjan Ghosh, Elena Musi, and Smaranda Muresan. 2020. Interpreting verbal irony: Linguistic strategies and the connection to theType of semantic incongruity. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 82–93, New York, New York. Association for Computational Linguistics.

Sayan Ghosh and Shashank Srivastava. 2022. ePiC: Employing proverbs in context as a benchmark for abstract language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3989–4004, Dublin, Ireland. Association for Computational Linguistics.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Souad Hamidi. 2023. The relationship between language, culture, and identity and their influence on one another. 3.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264, Dubrovnik, Croatia. Association for Computational Linguistics.

Risma Hayani. 2016. Figurative language on Maya Angelou selected poetries. *Script Journal: Journal of Linguistic and English Teaching*, 1:131.

Yusuke Ide, Joshua Tanner, Adam Nohejl, Jacob Hoffman, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. 2025. CoAM: Corpus of all-type multiword expressions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27004–27021, Vienna, Austria. Association for Computational Linguistics.

Hyewon Jang, Qi Yu, and Diego Frassinelli. 2023. Figurative language processing: A linguistically informed feature analysis of the behavior of language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9816–9832, Toronto, Canada. Association for Computational Linguistics.

Nevidu Jayatilleke and Nisansa de Silva. 2025a. Sidiac: Sinhala diachronic corpus. *arXiv preprint arXiv:2509.17912*.

Nevidu Jayatilleke and Nisansa de Silva. 2025b. Zero-shot OCR accuracy of low-resourced languages: A comparative analysis on Sinhala and Tamil. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 471–480, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Zhiying Jiang, Boliang Zhang, Lifu Huang, and Heng Ji. 2018. Chengyu cloze test. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 154–158, New Orleans, Louisiana. Association for Computational Linguistics.

Charles Jochim, Francesca Bonin, Roy Bar-Haim, and Noam Slonim. 2018. Slide - a sentiment lexicon of common idioms. In *International Conference on Language Resources and Evaluation*.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Aji, Genta Winata, Samuel Cahyawijaya, A. Aremu, Perez Ogayo, and Graham Neubig. 2023. Multilingual and multi-cultural figurative language understanding. pages 8269–8284.

Alina Karakanta, Mayra Nas, and Aletta G. Dorst. 2025. Metaphors in literary machine translation: Close but no cigar? In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 276–286, Geneva, Switzerland. European Association for Machine Translation.

Muhammad Farmal Khan and Mousumi Akter. 2025. Evaluating large language models on Urdu idiom translation. *Preprint*, arXiv:2510.17460.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. SemEval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47, Atlanta, Georgia, USA. Association for Computational Linguistics.

Tina Krennmayr and Gerard Steen. 2017. *VU Amsterdam Metaphor Corpus*, pages 1053–1071. Springer Netherlands, Dordrecht.

Murathan Kurfalı, Robert Östling, Johan Sjons, and Mats Wirén. 2020. A multi-word expression dataset for Swedish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4402–4409, Marseille, France. European Language Resources Association.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18601–18609.

Chaya Liebeskind and Yaakov HaCohen-Kerner. 2016. A lexical resource of Hebrew verb-noun multi-word expressions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 522–527, Portorož, Slovenia. European Language Resources Association (ELRA).

Changsheng Liu and Rebecca Hwa. 2017. Representations of context in recognizing the figurative and literal usages of idioms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31.

Thisiri Medagama. 2021. Idiomatic language complexities in translation with special reference to sinhalese and english. *Journal of Research in Humanities and Social Science*.

Wolfgang Mieder. 1997. Modern paremiology in retrospect and prospect. *Paremia*, 6:399–416.

Diego Moussallem, Mohamed Ahmed Sherif, Diego Esteves, Marcos Zampieri, and Axel-Cyrille Ngonga Ngomo. 2018. Lidioms: A multilingual linked idioms data set. *Preprint*, arXiv:1802.08148.

Andreas Musolff. 2017. Metaphor, irony and sarcasm in public discourse. *Journal of Pragmatics*, 109:95–104.

Farzana Nawaz, Tahira Jabeen, and Sadia Rather. 2025. The power of language and religious thoughts: A pragma-rhetorical analysis of israr ahmed's speech. *AGATHOS*, 16(2):167–182. Issue 31.

Gözde Özbal, Carlo Strapparava, and Serra Sinem Tekiroğlu. 2016. PROMETHEUS: A corpus of proverbs annotated with metaphors. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3787–3793, Portorož, Slovenia. European Language Resources Association (ELRA).

John Pavlopoulos, Panos Louridas, and Panagiotis Filos. 2024. Towards a Greek proverb atlas: Computational spatial exploration and attribution of Greek proverbs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11842–11854, Miami, Florida, USA. Association for Computational Linguistics.

Lotem Peled and Roi Reichart. 2017. Sarcasm SIGN: Interpreting sarcasm with sentiment based monolingual machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1690–1700, Vancouver, Canada. Association for Computational Linguistics.

Maria Pershina, Yifan He, and Ralph Grishman. 2015. Idiom paraphrases: Seventh heaven vs cloud nine. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 76–82, Lisbon, Portugal. Association for Computational Linguistics.

Ashmari Pramodya. 2023. Exploring low-resource neural machine translation for Sinhala-Tamil language pair. In *Proceedings of the 8th Student Research Workshop associated with the International Conference Recent Advances in Natural Language Processing*, pages 87–97, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

R. Priyanga, Surangika Ranathunga, and G. Dias. 2017. Sinhala word joiner. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 265–274, Kolkata, India. NLP Association of India.

Alexander Prochnow, Johannes E. Bendler, Caroline Lange, Foivos Ioannis Tzavellos, Bas Marco Göritzer, Marijn ten Thij, and Riza Batista-Navarro. 2024. IDEM: The IDioms with EMotions dataset for emotion recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8569–8579, Torino, Italia. ELRA and ICCL.

Carlos Ramisch, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, and Aline Villavicencio. 2016. How naked is the naked truth? a multilingual lexicon of nominal compound compositionality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–161, Berlin, Germany. Association for Computational Linguistics.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848, Online only. Association for Computational Linguistics.

John W. Ratcliff and David E. Metzener. 1988. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, 13(7):46–51.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Lok Regmi. 2015. Analysis and use of figures of speech. *Journal of NELTA Surkhet*, 4.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Richard M. Roberts and Roger J. Kreuz. 1994. Why do people use figurative language? *Psychological Science*, 5(3):159–163.

Yousef Sahari, Fawaz Qasem, Eisa Asiri, Ibrahim Alasmri, Ahmad Assiri, Shafi Alqahtani, and Hassan Mahdi. 2024. Evaluating the translation of figurative language: A comparative study of chatgpt and human translators. *CALR Linguistics Journal - Issue 15*.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, and 9 others. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Prateek Saxena and Soma Paul. 2020. EPIE dataset: A corpus for possible idiomatic expressions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4529–4536, Marseille, France. European Language Resources Association.

A.M. Senanayaka. 1880. *Athetha Wakya Deepanya*. Catholic Press.

John M. Senaveratna. 2005. *DictonarY of Proverbs of the Sinhalese*. Asian Educational Services, New Delhi, India.

Naziya Mahamdul Shaikh, Jyoti D. Pawar, and Mubarak Banu Sayed. 2024. Konidioms corpus: A dataset of idioms in Konkani language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9932–9940, Torino, Italia. ELRA and ICCL.

Dhirendra Singh, Sudha Bhingardive, and Pushpak Bhattacharyya. 2016. Multiword expressions dataset for Indian languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2331–2335, Portorož, Slovenia. European Language Resources Association (ELRA).

Nathalie Hau Sørensen, Sanni Nimb, Agnes Aggergaard Mikkelsen, and Jonas Jensen. 2025. The Danish idiom dataset: A collection of 1000 Danish idioms and

fixed expressions. In *Proceedings of the 1st Workshop on Nordic-Baltic Responsible Evaluation and Alignment of Language Models (NB-REAL 2025)*, pages 55–63, Tallinn, Estonia. The University of Tartu Library.

Caroline Sporleder, Linlin Li, Philip Gorinski, and Xaver Koch. 2010. Idioms in context: The IDIX corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Kenan Tang. 2022. Petci: A parallel English translation dataset of Chinese idioms. *Preprint*, arXiv:2202.09509.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.

Sarubi Thillainathan, Songchen Yuan, En-Shiun Annie Lee, Sanath Jayasena, and Surangika Ranathunga. 2025. Beyond vanilla fine-tuning: Leveraging multistage, multilingual, and domain-specific methods for low-resource machine translation. *Preprint*, arXiv:2503.22582.

Michael Toker, Oren Mishali, Ophir Münz-Manor, Benny Kimelfeld, and Yonatan Belinkov. 2024. A dataset for metaphor detection in early medieval Hebrew poetry. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 443–453, St. Julian's, Malta. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A large-scale Chinese IDiom dataset for cloze test. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 777–787, Florence, Italy. Association for Computational Linguistics.

Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48, Online. Association for Computational Linguistics.

A Existing Datasets Utilised

A.1 Germanic-Language Corpora

Sporleder et al. (2010) have introduced `IDIX` dataset which contains English idioms. In there, they have mentioned idioms as a contextual disambiguation problem. Rather than focusing on token-level labels, Haagsma et al. (2020) emphasises an inventory of potentially idiomatic expression types in English, that may be idiomatic depending on usage. The `PIE` dataset presented by Zhou et al. (2021) has been constructed to aid in the analysis of idiom paraphrasing by connecting idiomatic statements to alternatives that preserve meaning. `PIE` dataset by Adewumi et al. (2022), constructed from `BNC` and `UKWaC`, provides an additional comprehensive English-only structure where instances are labelled across different FoS, such as metaphor, simile, euphemism, and irony, alongside literal examples. This extends beyond a binary idiom/literal structure to facilitate fine-grained multi-class categorisation of figurative language.

The `VU Amsterdam Metaphor Corpus` by Krennmayr and Steen (2017) provides extensive manually annotated text that allows metaphor recognition in natural language for metaphor processing in English. It is frequently used to assess and train systems that need to recognise metaphorical usage on a large scale. Moreover, Saxena and Paul (2020) presented a more condensed English idiom-oriented dataset with an emphasis on modelling idiomatic phrases as evaluative targets. It is typically employed to determine whether representations capture the conventionalised meanings underlying idioms or address them compositionally. The

benchmark in Stowe et al. (2022) utilises paired instances to recast figurative understanding into a controlled evaluation format through the combination of a large semi-automatic section with a smaller manually selected gold set. Instead of focusing on the use of surface-level clues, it is meant to assess how effectively models handle figurative meaning, such as idioms and metaphors.

The dataset by Reddy et al. (2011) provides human judgments on the transparency of a compound's meaning and the strength of its components' contributions. This dataset serves as a common baseline for forecasting noun compounds' levels of (non-)compositionality. Liu and Hwa (2017) presents evaluation material for phrase-level robustness and rewriting where systems have to maintain meaning despite phrase replacements. This is helpful for evaluating phrase semantics and idiom-aware paraphrasing. In addition, `CoAM` by Ide et al. (2025) focuses on the behaviour of multiword expressions in English and supports identification studies in which *Multi-word Expressions* (MWEs) need to be regarded as single lexicalised components instead of distinct words. Furthermore, a number of English-only idiom benchmarks focus specifically on evaluating idiom competence rather than building linked lexical resources. Notable examples include `IDEM` by Prochnow et al. (2024), `IDIOMEM` by Haviv et al. (2023), and `SLIDE` by Jochim et al. (2018). With the objective to facilitate benchmarking and descriptive linguistic analysis in Danish, `The Danish Idiom Dataset` provides a selective collection of idioms and fixed expressions (Sørensen et al., 2025). Swedish resources enhance this idiom-specific focus by extending coverage to MWEs more broadly. This allows for wider-coverage modelling of formulaic language and provides annotated material for recognising lexicalised MWEs beyond idioms (Kurfalı et al., 2020). Furthermore, Germanic-language research frequently interacts with translation evaluation, especially in English-German contexts where specific idiom translation data allows for the methodical evaluation of MT/LLM errors such as literalization, semantic drift, and attenuation of figurative meaning during translation (Fadaee et al., 2018).

A.2 Indic-Language Corpora

`The Idiom Handling Dataset for Indian Languages` by Agrawal et al. (2018) provides idiom processing across several Indic languages such as Hindi, Urdu, Bengali, Tamil, Gujarati, Malayalam, Telugu, and typically includes mappings that enable cross-lingual handling, extending the coverage in Indic languages. In low-resource contexts, multilingual assessment and comparative analysis are enabled by (Agrawal et al., 2018).

In addition, the dataset presented by (Singh et al., 2016) focuses on Hindi and Marathi idioms/MWEs within Indic languages, offering annotated content for MWE/idiom recognition and mode ling in these languages. `Konidiom` by(Shaikh et al., 2024) provides idiom data for Konkani, a smaller, language-specific idiom resource that supports idiom research and resource development in a low-resource environment.

Khan and Akter (2025)'s dataset for Urdu focuses on translating idioms from Urdu and Roman Urdu, utilising idiom-focused test material to assess whether modern structures can preserve idiomatic meaning across script and language diversity. This is primarily an evaluation resource for translation behaviour under idiomaticity.

A.3 Romance-Language Corpora

Romance-language resources support a coherent discussion of how figurative meaning is represented within closely related languages and how well models transfer across them. By providing naturally grounded instances that allow idiom detection and interpretation in practical circumstances, `VIDiom-PT` supports this viewpoint for European Portuguese (Antunes et al., 2025). In contrast, `Prometheus` emphasises meaning recovery at the discourse level and is proverb-oriented, making it simpler to comprehend multilingual proverbs through English–Italian data. By allowing systematic comparison between related Romance languages, standardised multilingual assessment strengthens these language-specific techniques. `SemEval-2022 Task 2` provides a common benchmark for English, Portuguese, and Galician in similar language circumstances, allowing for controlled assessment of cross-lingual generalisation and transfer (Tayyar Madabushi et al., 2022).

A.4 Cross-Lingual Figurative Language Corpora

The large-scale cloze benchmark `ChID` by Zheng et al. (2019) is employed to evaluate idiom comprehension in Chinese resources. It requires models to select a suitable idiom to fill in a passage's blank. In addition to testing contextual idiom understanding through blank-filling. In addition to assessing

contextual idiom comprehending by blank-filling, the `Chengyu Cloze Test Dataset` by Jiang et al. (2018) emphasises semantic fit and discourse compatibility and delivers an invaluable, nearly equivalent evaluation environment.

Moreover, `PETCI` by (Tang, 2022) provides Chinese idioms related to English translations, facilitating the assessment of whether MT/LLM systems retain idiomatic meaning instead of generating literal, word-by-word renditions. Given this, it is extremely beneficial for controlled idiom translation testing. By enabling idiom identification as well as analysis in morphosyntactically rich contexts, where inflexion and flexible surface forms can complicate detection and interpretation, Slavic-language corpora expand figurative language study beyond English (Aharodnik et al., 2018; Donaj and Antloga, 2023). In order to allow both proverb retrieval/analysis and computational metaphor identification in a non-English setting, Greek corpora usually integrate structured proverb repositories with metaphor-annotated datasets (Pavlopoulos et al., 2024; Garcia et al., 2021). Through Hebrew and Arabic resources which facilitate MWE identification and metaphor detection in domain-specific contexts, including historically and stylistically unique texts that present additional model transfer challenges, Semitic corpora broaden coverage (Liebeskind and HaCohen-Kerner, 2016; Toker et al., 2024; Alsiyat et al., 2023).

As a way to improve cross-lingual mapping and interoperability, multilingual linked idiom resources represent idioms as interconnected lexical entities across languages and link them to external lexical-semantic inventories (Moussallem et al., 2018). Furthermore, multilingual shared benchmarks support systematic analysis of cross-lingual generalisation and provide consistent comparison of systems on MWEs, idiomaticity, and phrase-level semantics through providing standardised annotation guidelines and evaluation protocols across various languages (Savary et al., 2023; Korkontzelos et al., 2013; Tayyar Madabushi et al., 2022; Tedeschi et al., 2022). A summary of existing corpora, indicating the languages covered and the FoS addressed in the above studies, is shown in Table 6.

B Classification of Sinhalese Proverbs

Here we discuss the classification of Sinhalese proverbs based on different criteria as given below.

B.1 By the Nature of the Message (The Shape of the Message)

උපදේශ \ʊpʌðe:ʒʌ : Proverbs that contain a moral lesson or advice. While not all proverbs are adages, some are interchangeably used to provide direct guidance, such as "Don't burn your hand while the tongs are there".

විශ්වාස \vɪʒvɑ:sʌ: Proverbs that express a commonly accepted social truth or collective belief rather than a direct instruction. These are sometimes referred to as "Truth-principle proverbs" (Sathyadharma Pirulu). Examples include "A barking dog does not bite" or "Like eating the ear while sitting on the horn".

B.2 By Background Source (The Origin)

උපමා\ʊpʌmʌ : Ends in comparative markers. (වගේ*vʌge:*, සේ*se:*, මෙන්*mɛn* , වැනි*vænɪ*).

වාර්තා\vɑ:rθɑ:: Ends in hearsay markers (ලූ*lu:*).

ප්‍රශ්න\prʌʒnʌ : Ends in interrogative markers (ද*ðʌ*), often acting as rhetorical devices to prompt self-reflection (e.g.," සිත ඇත්නම් පත කුඩා ද? \ sɪθʌ æθnʌm pʌθʌ kʊdɑ: ðʌ?"

නිශේධ\nɪʒe:ðʰʌ (Negative): Ends in negation. (නැහැ*næhæ*, බෑ*bae:*, නෑ*nae:*).

B.3 By Grammatical Ending (The Marker)

කතා \ kʌθɑ: (Story-based): These proverbs rely on shared cultural memory. They are often unintelligible without knowledge of the specific folktale or historical event (e.g., "අන්දරේ සීනි කෑවා\ *ʌnðʌre: si:nɪ kæ:vɑ: vʌgɛj* " - Like Andare eating sugar).

ප්‍රපංච \ prpʌŋʧʌ (Phenomenon-based): These are derived from empirical observations of the agrarian environment, nature, or daily life (e.g., "පිණි දිය දැක නොතලන් නෙලා පලා*pɪnɪ ðɪjʌ ðækʌ nʊθʌlʌn nɛlɑ: pʌlɑ:* " - Do not crush the greens, seeing the dew).

සාහිත්‍ය \ sɑ:hɪθjʌ (Literature-based): These originate from classical texts such as the පන්සිය පනස් ජාතක\ *pʌnsɪjʌ pʌnʌs dʒɑ:θʌkʌ* or සුභාශිතය*subʰɑ:ʒɪθʌjʌ*, reflecting the influence of Buddhism and literacy on folk speech.

Among these, උපමා \ ʊpʌmʌ (Simile) sub-category is the most prevalent. This indicates that analogi-

Dataset	Languages	FOS Explored
IDIX (Sporleder et al., 2010)	English	Idioms
MAGPIE(Haagsma et al., 2020)	English	Potentially Idiomatic Expressions
PIE (Zhou et al., 2021)	English	Idiomatic Expressions (IE)
PIE(BNC and UKWaC) (Adewumi et al., 2022)	English	Metaphor, simile, euphemism, parallelism, personification, oxymoron, paradox, hyperbole, irony, and literal
MABL (Kabra et al., 2023)	Hindi, Indonesian, Javanese, Kannada, Sundanese, Swahili and Yoruba	Figurative language
VIDiom-PT (Antunes et al., 2025)	European Portuguese	Verbal Idioms
The Danish Idiom Dataset (Sørensen et al., 2025)	Danish	Idiomatic expressions and fixed expressions
LIDIOMS, DBnary,BabelNet (Moussallem et al., 2018)	English, German, Italian, Portuguese, and Russian	Idioms
Prometheus (Özbal et al., 2016)	English, Italian	Proverbs
VU Amsterdam Metaphor Corpus (Krennmayr and Steen, 2017)	English	Metaphors
MetaNet (Dodge et al., 2015)	English, Russian, Mexican Spanish, Iranian Farsi	Metaphors
EPIE (Saxena and Paul, 2020)	English	Idiomatic Expressions
IMPLI (Stowe et al., 2022)	English	Idiom, Metaphor
ePiC (Ghosh and Srivastava, 2022)	English	Proverbs
ChID (Zheng et al., 2019)	Chinese	Metaphor, Near-synonymy
UPD*(Reddy et al., 2011)	English	Compound Nouns
SemEval-2013 Task 5 Dataset (Korkontzelos et al., 2013)	English, German, Italian	Phrases
IdiomKB (Li et al., 2024)	English, Chinese, Japanese	Idioms
IDEM (Prochnow et al., 2024)	English	Idioms
IDIOMEM. (Haviv et al., 2023)	English	Idioms
ID10M (Tedeschi et al., 2022)	English, Chinese, Spanish, Dutch, French, German, Italian, Japanese, Polish, Portuguese	Idioms
PETCI (Tang, 2022)	Chinese, English	Idioms
AStitchInLanguageModels Dataset (Tayyar Madabushi et al., 2021)	English, Portuguese	Idioms
UPD* (Garcia et al., 2021)	English	Idioms
UPD* (Cordeiro et al., 2019)	English	Nominal Compounds
SLIDE (Jochim et al., 2018)	English	Idioms
Russian Idiom-Annotated Corpus (Aharodnik et al., 2018)	Russian	Idiom
UPD*(Fadaee et al., 2018)	English, German	Idioms,Idiom Translation Dataset
Idiom Handling Dataset for Indian Languages (Agrawal et al., 2018)	English, Hindi, Urdu, Bengali, Tamil, Gujarati, Malayalam, Telugu	Idioms
Chengyu Cloze Test Dataset (Jiang et al., 2018)	Chinese	Idioms
Multilingual Lexicon of Nominal Compound Compositionality (Ramisch et al., 2016)	English, French, Portuguese	Nominal Compounds
UPD* (Pershina et al., 2015)	English,Idioms	Idiom Paraphrase Dataset
Phrasal Substitution Dataset (Liu and Hwa, 2017)	English	Idiomatic Expressions
CoAM (Ide et al., 2025)	English	MWEs
ParaDiom (Donaj and Antloga, 2023)	Slovene, English	Idiomatic Texts
Konidioms Corpus (Shaikh et al., 2024)	Konkani	Idioms
Multi-word Expression Dataset for Swedish (Kurfalı et al., 2020)	Swedish	Multi-word Expression
PARSEME Corpus Release 1.3 (VMWEs) (Savary et al., 2023)	Arabic, Bulgarian, Chinese, Croatian, Greek, Hebrew, Hindi, Irish, Latvian, Lithuanian, Maltese, Slovene, Turkish	Idioms, multiword expressions (verbal MWEs)
SemEval-2022 Task 2 Dataset (Tayyar Madabushi et al., 2022)	English, Portuguese, Galician	Idioms
UPD*(Singh et al., 2016)	Hindi, Marathi	Idioms, MWEs
UPD* (Liebeskind and HaCohen-Kerner, 2016)	Hebrew	MWEs (incl. idiom-like fixed expressions)
Greek Proverb Atlas(Pavlopoulos et al., 2024)	Greek	Proverbs
UPD* (Florou et al., 2018)	Greek	Metaphor
UPD* (Toker et al., 2024)	Hebrew	Metaphor
UPD* (Khan and Akter, 2025)	Urdu, Roman Urdu	Idioms
AMC (Alsiyat et al., 2023)	Arabic	Metaphor

Table 6: Existing Datasets Summary. *Corpora named 'UPD' represent the *Unnamed Primary Dataset(s)*, which includes papers that have released/utilised datasets without specific names.

cal reasoning, understanding one concept in terms of another, is the primary cognitive tool used in Sinhala folk wisdom. වාර්තා \vaːrθaː (Report) category is the second most common proverb structure. The prevalence of the particle ලු \ *luː* (it is said) underscores the importance of oral tradition and collective knowledge in Sri Lankan culture, wisdom is validated not by the speaker's authority, but by the fact that "it has been said" by ancestors.

C Sinhala Proverbs vs Sinhala Idioms

The Dichotomy of Sinhala Proverbs and Sinhala Idioms: While both categories function as figurative devices, they are distinguishable through three primary dimensions: Syntactic Structure, Semantic Deductibility, and Pragmatic Function.

Sinhala (සිංහල \ sɪŋhʌlʌ) : ඉත්තෑවගේ ගුලේ කබල්ලෑව වැදී "මුත්තාපා කීවත් යන්නේ නෑ" කිව්වාලු \ ɪθθæːvʌgeː guleː kʌbʌlːlæːvʌ vʌðiː muθθɑːpɑː kiːvʌθ jʌnneː næː kɪvvɑːlʊ

Type of FOS : [විශ්වාස][කතා][උපමා] \ [vɪʒvɑːsʌ][kʌθɑː][ʊpʌmɑː]

Literal / visual image : The ant-eater, who forcibly occupied the porcupine's hole, swore by his forbears that he would never leave it.

Corresponding FOS in English : Possession is nine-tenths of the law.

What it really implies : Taking possession of other people's property through deceit.

Additional Context : In a certain forest, a porcupine lived inside a cave. One day, an anteater, who had lost his way while traveling, came across this cave and asked the porcupine if he could stay there for shelter. The porcupine kindly agreed and allowed the anteater to stay inside. However, the next morning, the anteater showed no sign of leaving. Since the cave was too small for both of them to live together, the porcupine politely requested the anteater to leave. "If you don't like it, then you can go and find another place. I'm quite comfortable here," the anteater replied. Angered, the porcupine raised his sharp quills and attacked the anteater. But the anteater's body was covered in thick, coarse hide, so the porcupine's blows had no effect. The anteater remained in the cave, while the porcupine was forced to leave and find another shelter.

Figure 4: An example of a record on the dataset.

Semantic Deductibility (Opacity vs. Transparency): Idioms in Sinhala often exhibit high semantic opacity; a learner cannot easily deduce that " කහ \ kʌhʌ" in " ඌරන්ට කහ \ uːrʌntʌ kʌhʌ" implies "wasting resources." However, Proverbs are often semantically translucent. Even a first-time listener can deduce the meaning of " ගහ දන්න අයට කොළ කඩා පානවා \ gʌhʌ ðʌnnʌ ʌjʌtʌ kɒlʌ kʌdɑː pɑːnʌvɑː" (Showing leaves to those who know the tree) based on the imagery of deception and expertise.

Pragmatic Function: ප්‍රස්තාපිරුළු \ prʌsθapɪrulu are didactic; they convey general truths, social beliefs, or moral advice (උපදේශ \ ʊpʌðeːʒʌ). වාග්සම්ප්‍රදා \ vɑːg sʌmprʌðɑː are descriptive; they categorise a state of being or an action without necessarily offering a moral judgment.

Dominance of Idioms: වාග්සම්ප්‍රදා \ vɑːg sʌmprʌðɑː constitute the overwhelming majority of the dataset. This quantitative dominance suggests that Sinhala speakers prioritise "descriptive efficiency" in daily language, using short, culturally loaded phrases to quickly describe complex situations, over the more formal, structured wisdom of proverbs.

D Dataset Annotation

The dataset was annotated by filling in the fields. Not all fields were filled in for all records, as shown in Table 1. Figure 4 contains an example of a record in the dataset.

E Cultural Analysis

The Source Domain explores the abstract imagery and objects used in the FoS to deliver the message, whilst the target theme is used to identify the messages delivered by the various FoS.

E.1 Specific Cultural Codes

Certain symbols carry specific, unchangeable meanings in the Sinhala cultural lexicon. The following are some of the examples utilised in Sinhala FoS.

The Elephant (Power & Scale): The elephant is the cultural yardstick for greatness. It is used to contrast "the great" with "the small." It represents forces that are often too big to manage or criticise.

The Dog (Low Status): In contrast to the elephant, the dog is consistently used to represent unworthiness or low social status. It serves as a warning of what happens when one lacks dignity.

The Tree (Character): Trees are almost always metaphors for moral character. A person is judged like a tree, by their "fruit" (utility to society) or their "wood" (strength/weakness).

E.2 Emotional Landscape

The sentiment analysis shows that the vast majority of FoS (83% of the data) are *Neutral*. They are not optimistic or pessimistic; they are descriptive. The culture does not say "Life is good" or "Life is bad"; it says, "If you take this action, the corresponding outcome will occur inevitably." It values truth over comfort.

Table 7 provides a comprehensive overview of the cultural analysis, summarising the frequency of literal imagery and the specific thematic domains explored within the SINFOS dataset.

Category	No. of Occurrences
Source Domain (Literal Imagery)	
Body & Senses (Somatic)	242
Nature & Agriculture	194
Animals (Fauna)	148
Household & Daily Life	144
Target Theme (Cultural Meaning)	
Ethics & Moral Character	162
Karma & Consequence	127
Impermanence & Uncertainty	121
Social Status & Hierarchy	73
Human Relations & Conflict	64

Table 7: Distribution of literal source domains and abstract cultural themes observed in the SINFOS dataset via hybrid thematic analysis.

F Model Classification

The results of classifying proverbs and idioms are summarised in Table 8. Word2Vec showed the best performance for Naive-Bayes and Linear SVC in terms of recall and accuracy. In contrast, TF-IDF 3-gram vectorisation excelled with Random Forest, XGBoost, and the ensemble model combining these with Linear SVC.

Experiments	Acc.	P-Rec.	I-Rec.
TF-IDF (Unigram) Multimodal Naive-Bayes	0.7167	0.59	0.81
TF-IDF (Char 3-gram) Multimodal Naive-Bayes	0.8348	0.79	0.87
Gaussian Naive-Bayes Word2Vec	0.9034	0.92	0.89
TF-IDF (Unigram) Linear SVC	0.7639	0.64	0.86
TF-IDF (Char 3-gram) Linear SVC	0.8712	0.81	0.92
Linear SVC Word2Vec	0.9034	0.90	0.91
Random Forest (tuning) TF-IDF (Unigram)	0.8026	0.66	0.91
Random Forest (tuning) TF-IDF (Char 3-gram)	0.8927	0.83	0.94
Random Forest (tuning) Word2Vec	0.8755	0.81	0.93
XGBoost (tuning) TF-IDF (Unigram)	0.8090	0.65	0.93
XGBoost (tuning) TF-IDF (Char 3-gram)	0.9013	0.86	0.93
XGBoost (tuning) Word2Vec	0.8690	0.83	0.90
TF-IDF (Char 3-gram) Voting Ensemble	0.9056	0.85	0.95
Voting Ensemble Word2Vec	0.8884	0.85	0.92
Bi-LSTM	0.9163	0.86	0.96
Deep NN	0.9270	0.94	0.92

Table 8: Model Performance: Accuracy, Proverbs Recall (P-Rec.), and Idioms Recall (I-Rec.).

G Performance of all LLMs

A brief overview of each metric's blind spots and how each metric mitigates the other's is provided in Table 9.

Metric	Blind Spot	Mitigation Strategy
Cosine Similarity	The "Keyword Bag" Problem: the model may achieve high scores by guessing relevant keywords even if the grammatical structure is flawed.	Fidelity acts as a "Logic Gate," requiring semantic validity rather than just keyword overlap.
Fidelity Score	The "Hyper-Literal" Problem: creative paraphrases with different structures might be penalised.	Cosine Similarity permits creative phrasing; high similarity with low fidelity suggests a valid non-standard translation.

Table 9: Evaluation Metrics and Mitigation of their Blind Spots

The Figures 5 and 6 represent the Cosine Similarity scores and Fidelity Scores of all the models across seven different categories. Along with ආප්තෝපදේශ \ ɑːpθɔːpʌðeːʒʌ, the models seem to have decent performances for proverbs associated with nature as they seem to be able to decipher the meaning using the phenomenon. In the case of කියමන් \ kɪjʌmʌn though, as in proverbs based on folklore, the language models seem to struggle. This is tied with the fact that unlike ආප්තෝපදේශ \ ɑːpθɔːpʌðeːʒʌ, කියමන් \ kɪjʌmʌn are more specific to the language.

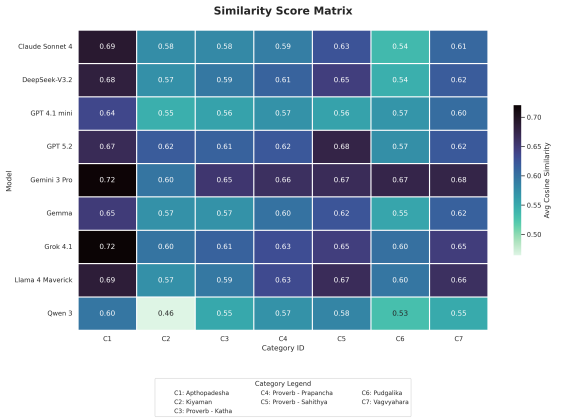


Figure 5: Benchmarking LLM Performance: Cosine Similarity.

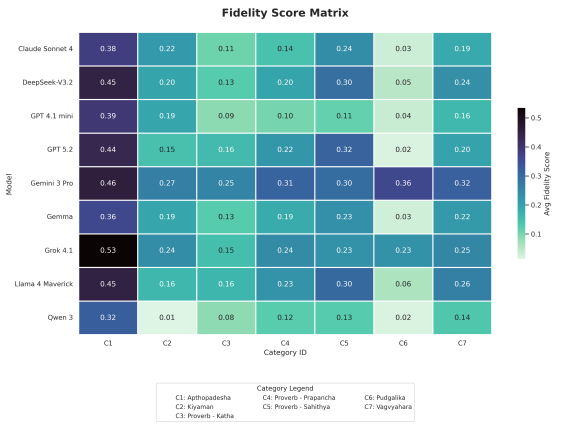


Figure 6: Benchmarking LLM Performance: Fidelity Scores.