A Parallel Cross-Lingual Benchmark for Multimodal Idiomaticity Understanding

TORUNOĞLU-SELAMET, Dilara, ARSLAN, Doğukan, WILKENS, Rodrigo, HE, Wei, ERYIĞIT, Doruk, PICKARD, Thomas, PAGANO, Adriana S, VILLAVICENCIO, Aline, ERYIĞIT, Gülşen, ABUCZKI, Ágnes, CARDOSO, Aida, LAZARENKA, Alesia, ALMASSOVA, Dina, MENDES, Amália, KANELLOPOULOU, Anna, BROSA-RODRIGUEZ, Antoni, VALKOVSKA, Baiba, WOJTOWICZ, Beata, PEDERSEN, Bolette, HIDALGO-TERNERO, Carlos Manuel, LIEBESKIND, Chaya, JOKIĆ, Danka, ALVES, Diego, TRIANTAFYLLIDI, Eleni, VELLDAL, Erik, PHILIPPY, Fred, OLESKEVICIENE, Giedre Valunaite, RIZGELIENE, Ieva, SKADINA, Inguna, LOBZHANIDZE, Irina, HAUGEN, Isabell Stinessen, KRITO, Jauza Akbar, MARKOVIĆ, Jelena M, MONTI, Johanna, SAUCA, Josue Alejandro, DOBROVOLJC, Kaja, UGWUANYI, Kingsley O, RITUMA, Laura, ØVRELID, Lilja, AGRO, Maha Tufail, ABJALOVA, Manzura, CHATZIGRIGORIOU, Maria, RAMOS, María del Mar Sánchez, PENDEVSKA, Marija, SEYYEDREZAEI, Masoumeh, SHAMSFARD, Mehrnoush, AHSAN, Momina, KHAN, Muhammad Ahsan Riaz, NORMAN, Nathalie Carmen Hau, AYYILDIZ, Nilay Erdem, HOSSEINI-KIVANANI, Nina, LIGETI-NAGY, Noémi, NAEEM, Numaan, KANISHCHEVA, Olha, YATSYSHYNA, Olha, OREL, Daniil, GIOMMARELLI, Petra, OSENOVA, Petya, GARABIK, Radovan, SEMOU, Regina E, REBECHI, Rozane, PRANIDA, Salsabila Zahirah, TOUILEB, Samia, NIMB, Sanni, AHMAD, Sarfraz, SHARIPOVA, Sarvinoz, GOLAN, Shahar, JI, Shaoxiong, ABOH, Sopuruchi Christian, SUCUR, Srdjan, MARKANTONATOU, Stella, OLSEN, Sussi, TAJALLI, Vahide, LIPP, Veronika, GIOULI, Voula, ERAYDIN, Yelda Yeşildal, SAABERI, Zahra and XIE, Zhuohan

A Parallel Cross-Lingual Benchmark for Multimodal Idiomaticity Understanding

Dilara Torunoğlu-Selamet[*,1,†], Doğukan Arslan[*,1], Rodrigo Wilkens[*,2], Wei He[*,2], Doruk Eryiğit, Thomas Pickard[*,3], Adriana S. Pagano[*,4], Aline Villavicencio[*,2,3], Gülşen Eryiğit[*,1], Ágnes Abuczki, Aida Cardoso, Alesia Lazarenka, Dina Almassova, Amalia Mendes, Anna Kanellopoulou, Antoni Brosa-Rodríguez, Baiba Valkovska, Beata Wojtowicz, Bolette Pedersen, Carlos Manuel Hidalgo-Ternero, Chaya Liebeskind, Danka Jokić, Diego Alves, Eleni Triantafyllidi, Erik Velldal, Fred Philippy, Giedre Valunaite Oleskeviciene, Ieva Rizgeliene, Inguna Skadina, Irina Lobzhanidze, Isabell Stinessen Haugen, Jauza Akbar Krito, Jelena M. Marković, Johanna Monti, Josue Alejandro Sauca, Kaja Dobrovoljc, Kingsley O. Ugwuanyi, Laura Rituma, Lilja Øvrelid, Maha Tufail Agro, Manzura Abjalova, Maria Chatzigrigoriou, María del Mar Sánchez Ramos, Marija Pendevska, Masoumeh Seyyedrezaei, Mehrnoush Shamsfard, Momina Ahsan, Muhammad Ahsan Riaz Khan, Nathalie Carmen Hau Norman, Nilay Erdem Ayyıldız, Nina Hosseini-Kivanani, Noémi Ligeti-Nagy, Numaan Naeem, Olha Kanishcheva, Olha Yatsyshyna, Daniil Orel, Petra Giommarelli, Petya Osenova, Radovan Garabik, Regina E. Semou, Rozane Rebechi, Salsabila Zahirah Pranida, Samia Touileb, Sanni Nimb, Sarfraz Ahmad, Sarvinoz Sharipova, Shahar Golan, Shaoxiong Ji, Sopuruchi Christian Aboh, Srdjan Sucur, Stella Markantonatou, Sussi Olsen, Vahide Tajalli, Veronika Lipp, Voula Giouli, Yelda Yeşildal Eraydın, Zahra Saaberi, Zhuohan Xie

[*] Core authors [1] Istanbul Technical University [2] University of Exeter
[3] University of Sheffield [4] Federal University of Minas Gerais

Abstract

Potentially idiomatic expressions (PIEs) carry meanings inherently tied to the everyday experience of a given language community. As such, they constitute an interesting challenge for assessing the linguistic (and to some extent cultural) capabilities of NLP systems. In this paper, we present XMPIE, a parallel multilingual and multimodal dataset of potentially idiomatic expressions. The dataset, containing 34 languages and over ten thousand items, allows comparative analyses of idiomatic patterns among language-specific realisations and preferences in order to gather insights about shared cultural aspects. This parallel dataset allows evaluation of language model performance for a given PIE in different languages and whether idiomatic understanding in one language can be transferred to another. Moreover, the dataset supports the study of PIEs across textual and visual modalities, to measure to what extent PIE understanding in one modality transfers or implies in understanding in another modality (text vs. image). The data was created by language experts, with both textual and visual components crafted under multilingual guidelines, and each PIE is accompanied by five images representing a spectrum from idiomatic to literal meanings, including semantically related and random distractors. The result is a high-quality benchmark for evaluating multilingual and multimodal idiomatic language understanding.

Keywords: Multiword Expressions, Machine Translation, Multilingual models, Multimodal models

1. Introduction

As a widely studied class of non-compositional multiword expressions, idioms such as “green fingers” and “kick the bucket”, pose persistent challenges for both humans and natural language processing (NLP) systems (Sag et al., 2002). Due to the cultural knowledge and shared conceptualisations they embody, idioms can be challenging not only for (non-native) speakers (Charteris-Black, 2002; Kovecses, 2010), but also in tasks like machine translation (Sag et al., 2002). Indeed, idioms bring to light a range of cross-linguistic dynamics. On the one hand, speakers of closely related languages often share similar idiomatic realisations, while on the other hand, idioms may be specific to a linguistic community and lexically non-transferable and opaque, leading to entirely differ-

[†]Corresponding author: torunoglud@itu.edu.tr

ent figurative mappings (Irujo, 1986). For example, "bad apple" is directly transferable into Turkish as "çürük elma" (lit. "rotten apple"), whereas the English idiom "bear market" has no idiomatic equivalent and is instead paraphrased descriptively (e.g., "düşen piyasa" lit. "declining market"), while the Brazilian Portuguese "levar uma bola nas costas" (lit. "take a ball on the back"), conveying the idea of being betrayed, may be understood by a Turkish speaker given the analogous "sırtından bıçaklanmak" ("to take a knife in the back"), with a similar metaphorical frame of unexpected betrayal. This cultural and cross-lingual dimension positions idioms as an ideal testbed for investigating linguistic diversity and the capacity of NLP models to generalize underlying meaning across languages.

There has been a growing emphasis on cultural benchmarks (Chiu et al., 2025; Khanuja et al., 2024; Romero et al., 2024) when evaluating large language models (LLMs), reflecting the need to assess, not only their general linguistic competence, but also their sensitivity to diverse cultural contexts and knowledge. Since idioms are deeply embedded in cultural knowledge and often resist literal translation, the introduction of a multimodal parallel dataset of potentially idiomatic expressions (PIEs) provides significant value for facilitating crosslingual intercultural comparison and systematically evaluating multilingual representativeness in LLMs. Moreover, PIEs are an interesting challenge due to the data-intensive nature of LLMs, which creates a serious obstacle to building accurate representation for low-frequency and long-tail phenomena, like PIEs, especially in multilingual and low resource scenarios.

While contemporary LLMs achieve strong performance on a variety of NLP tasks (Zubiaga, 2024), they often struggle with idiomatic expressions, highlighting a gap in their capacity for idiomatic understanding (Phelps et al., 2024; Arslan et al., 2025). Traditional idiom processing benchmarks, which typically frame idiom comprehension as a classification task, have been criticized as not fully reflective of a model's grasp of idiomatic meaning (Boisson et al., 2023; He et al., 2025), particularly given that general classification performance of LLMs is still limited. Alternatively, paraphrasing and multimodal approaches have been proposed to better assess idiomatic competence, with recent datasets integrating visual modalities alongside text to challenge models in multimodal idiom understanding (Pickard et al., 2025).

This paper introduces a benchmark that paves the way for the evaluation of the extent to which LLMs display accurate idiom comprehension across languages and across modalities. The *Cross-lingual and Multimodal Potentially Idiomatic Expressions* (XMPIE) dataset, a parallel cross-lingual benchmark for multimodal idiomaticity understanding. XMPIE is inspired by AdMIRe 1.0 (Pickard et al., 2025) which covers 2 languages (English and Portuguese), and extends it to 34 languages, ranging from widely spoken ones to those endangered either sociolinguistically (viz., Aromanian and Luxembourgish) or digitally (viz., Igbo, Catalan, Greek, Latvian, and Lithuanian) according to UNESCO (Moseley, 2010) and the European Language Equality (ELE) report (Ananiadou et al., 2012). Starting from seed English PIEs and extending them to these languages XMPIE has 3054 PIEs and 7040 images in total for figurative and literal meanings, along with distractors (Figure 1). In the paper, we discuss the construction of this parallel cross-linguistic dataset for understanding multimodal idiomaticity and the analyses carried out on this. The paper starts with related work in §2, and the annotation methodology in §3 while §4 presents the resulting dataset, §5 the benchmark evaluation, and §6 the conclusions.

2. Related Work

Understanding how computational models represent and predict compositional meaning relies notably on the development of benchmark datasets. These resources have advanced, from early, monolingual collections with compositionality ratings (Cook et al., 2008) to more complex multilingual corpora that capture nuanced, context-dependent phenomena (Haagsma et al., 2020). In parallel with advancements in human-annotated data, recent work has also begun to explore innovative strategies for corpus creation and the integration of multimodal evaluation items.

A summary of these datasets is shown in Table 1, where a prevalence of European languages can be seen. Despite advances in scale and diversity, the majority of existing idiom datasets are not constructed in a parallel fashion, meaning that even in multilingual datasets, idiomatic expressions are not systematically aligned across languages. As a result, while these corpora enable monolingual and, to some extent, multilingual evaluation, they do not facilitate direct cross-linguistic comparison of idiom usage or support fine-grained investigations of how idiomatic meaning is preserved or altered across languages. This lack of parallelism remains a critical gap, particularly for research on multilingual representation learning and cross-lingual transfer.

Foundational work in this area was primarily monolingual: from Reddy et al. (2011) English noun compound dataset, which established a strong link between literal word meaning and overall phrase compositionality, and were extended by Cordeiro et al. (2019) with human compositional-

(a) idiomatic (b) idiomatic-related (c) literal-related (d) literal (e) distractor

Figure 1: Image data for *"green fingers"* idiomatically denoting *skill in gardening.*

Dataset	#Size	#Idioms	Language
VNC-Tokens (Cook et al., 2008)	2,566	53	en
Open-MWE (Hashimoto and Kawahara, 2009)	102,856	146	ja
Sporleder and Li (Sporleder and Li, 2009)	3,964	17	en
IDIX (Sporleder et al., 2010)	5,836	78	en
SemEval-2013 Task 5b (Korkontzelos et al., 2013)	4,350	65	en
PARSEME (Savary et al., 2015)	274,376	13,755	bg, cs, fr, de, he, it, lt, mt, el, pl, pt, ro, sl, es, sv, tr
MAGPIE (Haagsma et al., 2020)	56,622	2,007	en
EPIE (Saxena and Paul, 2020)	25,206	717	en
AStitchInLang.Models (Tayyar Madabushi et al., 2021)	6,430	336	en, pt
ID10M$_{gold}$ (Tedeschi et al., 2022)	800	470	de, en, es, it
ID10M$_{silver}$ (Tedeschi et al., 2022)	262,781	10,118	de, en, es, fr, it, ja, nl, pl, pt, zh
SemEval-2022 Task 2 (Tayyar Madabushi et al., 2022a)	8,683	50	en, gl, pt
Dodiom (Eryiğit et al., 2022)	12,706	73	it, tr
idiom-corpus-llm (Arslan et al., 2025)	34,600	173	en, ja, it, tr

Table 1: A summary of various idiom corpora, detailing the number of sentences, the number of idioms, and the languages included for each (Arslan et al., 2025).

ity judgments for nominal compounds in English, French, and Portuguese, supporting multilingual evaluation. Other multilingual resources include SemEval-2022 Task 2 (Tayyar Madabushi et al., 2022b, 2021) providing 8,683 multilingual entries for idiomaticity detection and sentence representation for English, Portuguese and Galician. MultiCoPIE (Sentsova et al., 2025) is a multilingual corpus of PIEs in Catalan, Italian, and Russian, with annotations designed to analyze factors like lexical overlap in cross-lingual transfer learning.

Beyond increasing language coverage, research has also focused on capturing more granular, context-dependent aspects of idiomaticity. For instance, He et al. (2025) introduced a large-scale dataset for English and Portuguese containing minimal pairs, human judgments at both type and token levels, paraphrases, and contextual occurrences. Another example is DICE (Mi et al., 2025), a contrastive dataset designed specifically to test whether LLMs can effectively use context to disambiguate idiomatic expressions and to systematically analyze their limitations.

A more recent and challenging frontier is the extension of this research into multimodal settings, exploring how idiomatic meaning is conveyed through both text and images. The Image Recognition of Figurative Language (IRFL) dataset from Yosef et al. (2023), which established that state-of-the-art vision-language models significantly underperform humans on multimodal tasks involving metaphors, similes, and idioms. Building on this, Saakyan et al. (2025) developed V-FLUTE, a dataset that adds a layer of explainability by requiring models to generate textual justifications for their visual entailment decisions on figurative language. The scope of multimodal research has also expanded to specific languages and domains. For instance, Wang et al. (2025) created MChIRC, a large-scale dataset specifically for Chinese idiom comprehension, while Tong et al. (2025) compiled the HUMMUS dataset to analyze the interplay of humorous metaphor and idioms.

Parallel to the development of these annotated corpora, recent studies have explored innovative strategies for data creation to overcome the costs of manual annotation. One direction involves novel human-in-the-loop methods, such as the gamified crowdsourcing framework introduced by Eryiğit et al. (2022) to engage native speakers in creating and rating idiom examples. In a complementary, model-centric approach, Arslan et al. (2025) investigated using LLMs themselves to generate synthetic idiom corpora, offering a potentially

more scalable and efficient alternative to human-annotated datasets.

However, a massively cross-lingually aligned resource for exploring idiomaticity in both language production and model evaluation is still missing. Therefore, in this paper, we address this and present a parallel dataset across multiple languages and modalities and discuss the protocol adopted to facilitate annotation and coordination.

3. Annotation Methodology

Starting from a seed set of PIEs in English, 78 native or highly fluent language experts collaboratively identified equivalent idioms in their respective language variants and generated relevant images depicting both figurative and literal meanings along with distractors using image generation systems.

Language experts were recruited through open invitation through UniDive (Savary et al., 2024) and participated in three online workshops, where they were presented with step-by-step instructions about the annotation process. They were also given written guidelines while additional consultation was offered upon demand.

For each seed English PIE (e.g., "bad apple"), annotators provided: (1) a literal word-by-word translation into their language (e.g., Ukrainian "gnile âbluko"); (2) a transliterated version of the literal translation (when applicable) ("гниле яблуко"); (3) the idiomatic equivalent of the PIE in their language ("paršiva vìvcâ"); (4) its literal (word-by-word) English translation[1] ("lousy sheep") and (5) a transliterated version of that idiomatic form (when applicable) ("паршива вівця").

The second step was to generate images. To do this, we used Discord[2] to facilitate communication and sharing during the annotation process from a centralized point and to enable automatic collection of image generation prompts. Each language was assigned its own Discord channel for the language experts to collaborate in preparing the PIE images and to conduct the necessary discussions. Midjourney[3] was adopted for image generation and each language was provided with a one-month Midjourney subscription. The administrators were members of all channels, monitored the process, and responded to annotators' questions as needed.

For PIE image generation, annotators produced *five* candidate images that range from the idiomatic to the literal meaning (Figure 1): (1) an image of the *idiomatic* sense, (2) an image for an *idiomatic–related* sense, (3) an image with a *literal–related* sense, (4) one for the *literal* sense, and (5) one for an *unrelated* distractor. This provides an image representation for the idiomatic and literal meanings of a PIE as well as easy and difficult distractors.

These items are included in in order to emphasise the importance of processing the **semantics** of the image contents over surface elements in correctly identifying the most appropriate representation of a given expression in context and avoid potential shortcuts (such as selecting the only image containing a person when one is mentioned in the context sentence).

Textual prompts providing descriptions for each of the five target senses were used to generate images. Language experts identified the two target senses (idiomatic and literal) before generating the three distractors. Where possible, the semantically related distractors stayed in the same broad semantic category as the target (e.g., an object for object-denoting PIE). For generating the semantically related distractors, strategies included focusing on the individual words or on aspects of the specific senses (e.g. "a bag of apples" for "bad apple" and "a shelf with small pieces of cheese" for "big cheese"). One of the challenges is that prompts need to be framed in terms of concrete and visually grounded descriptions ("recipe book" rather than the more abstract "instructions") and avoiding potential issues that image generators struggle with (e.g., legible text on signs; fine-grained hand poses; subtle emotions). Abstract adjectives were conveyed via visual proxies (e.g., for an important person the prompt would include a figure of authority). When a figurative sense is hard to portray directly, a literal-looking scene that cues the idiomatic reading (e.g., a reviewer or judge for "armchair critic") would be used. During training, "good" and "bad" prompt examples were provided to the annotators for calibration. Image prompts were executed in private ("stealth") mode via direct messages with the Midjourney bot to prevent public leakage of the images. Final images were saved as high-resolution PNGs named 1–5 for idiomatic, idiomatic-related, literal-related, literal, and random distractor types.

4. The XMPIE Dataset

The dataset includes data for 34 language variants, viz., Aromanian, Azerbaijani, Bulgarian, Catalan, Chinese, Danish, Farsi, Georgian, Greek, Hebrew, Hungarian, Igbo, Indonesian, Italian, Javanese, Kazakh, Latvian, Lithuanian, Luxembourgish, Macedonian, Norwegian, Brazilian and European Portuguese, Russian, Serbian, Slo-

[1] We included literal translation of the idiomatic equivalent back into English for facilitating analyses and enabling cross-linguistic semantic transparency.

[2] https://discord.com/

[3] https://www.midjourney.com/

vak, Slovenian, Ecuadorian and European Spanish, Swahili, Turkish, Ukrainian, Urdu, Uzbek with contributions from 78 language experts, producing over 3054 expressions from the English seed PIEs. For a subset of those, 5 images were generated resulting in a total of 7040 images.

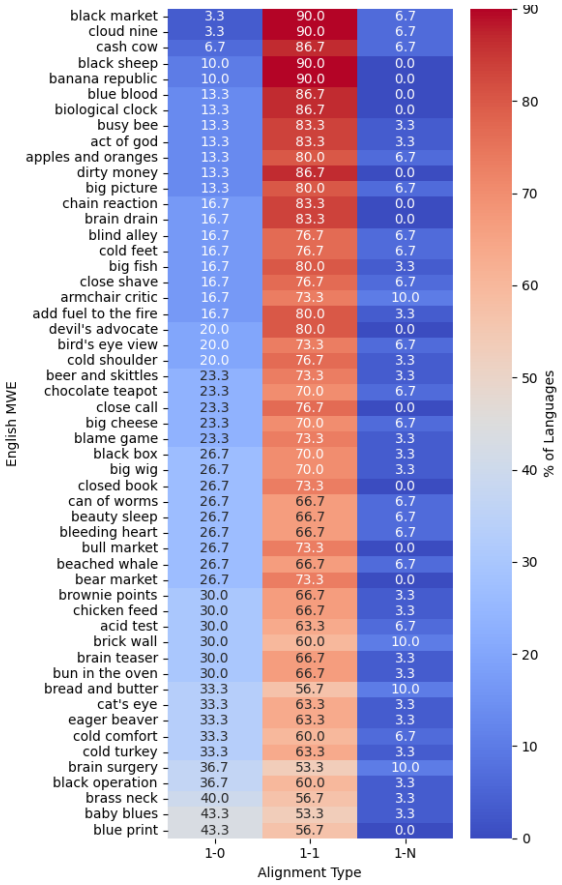


Figure 2: Percentage distribution of alignment types (*1-0*, *1-1*, *1-N*) across PIEs. Only idioms in more than 20 languages shown.

While all 34 language variants in the dataset underwent rigorous cross-checking and secondary human review, they currently exhibit varying degrees of coverage and completeness, reflecting their different stages of data collection. Consequently, we categorize the languages into two dataset versions according to the extent and completeness of their coverage:

- Core Dataset: Languages in this dataset contain all required potentially idiomatic expressions and modalities, meeting all strict quality thresholds during reviews. This dataset includes 21 language variants: Azerbaijani, Bulgarian, Chinese, Georgian, Greek, Igbo, Italian, Kazakh, Lithuanian, Norwegian, Portuguese (Brazilian and European), Russian, Serbian, Slovak, Slovenian, Spanish (Ecuadorian and European), Turkish, Ukrainian, and Uzbek.
- Extended Dataset: For some languages, data collection is still ongoing and, at the time of publication, some elements may still be missing (such as context sentences for specific PIEs) or may require additional rounds of review. Nevertheless, these languages are included because of their substantial value for linguistic diversity and representation, enabling broader cross-lingual analysis. The Extended Dataset comprises 13 language variants: Aromanian, Catalan, Danish, Farsi, Hebrew, Hungarian, Indonesian, Javanese, Latvian, Luxembourgish, Macedonian, Swahili, and Urdu.

Across all languages, leaders reported structural, semantic, and cultural specificities (Table 2). Although several idioms had an identical or near identical counterpart (1-1 in Table 2), such as “bad apple”, others implicated structural differences, as is the case of noun phrases such as “close shave” having as counterparts adverbial phrases, such as “o vlások” (‘by a hair’, Slovak). Others lacked direct lexical or idiomatic equivalents (1-0 in Table 2), reflecting cultural or conceptual gaps that required paraphrasing or descriptive expressions. Even when equivalents existed, partial semantic mismatches were frequent, as figurative scope, tone, or emotional polarity rarely aligned perfectly across languages. Cultural imagery also played a role in finding suitable idiomatic equivalents, as metaphors grounded in local experience are often needed to replace or revise the original image generated for English. Some idioms became lexicalized into single words (1-0 in Table 2), while others expanded into multiple variants or broader idioms with overlapping meanings (1-N in Table 2). The influence of English was also reported through calques and loan translations.

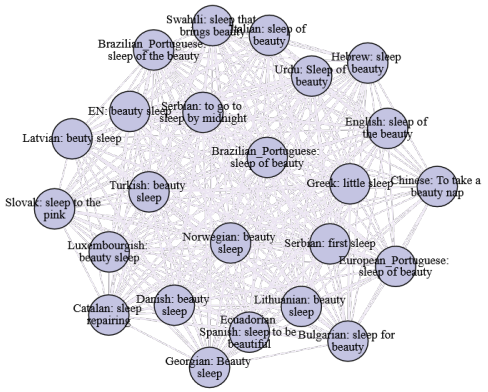
Based on the qualitative reports by the language experts, distinct types of idiom mappings were identified, which characterise how idiomatic meanings differ across languages, in relation to the English seed PIE, as summarized in Table 2. These fine-grained mappings can be grouped into four rough-grained types of alignment with the English seed PIEs: 1-1, when the idiomatic expression is also realised idiomatically; 1-0, when the meaning is conveyed but not as an expression; 1-N, when multiple PIEs are available, possibly reflecting ambiguity or variation; and N-1, when multiple English PIEs are equivalent. Figure 2 shows the proportion of a subset of types across English PIEs. Focusing on the English PIEs which have an idiomatic parallel in most other languages, these include cases like “black market” or “cash cow” which have correspondences across 90% and 87% of the languages, reflecting their conceptual salience and likely metaphorical shared

Alignment Type	Nature of Mapping	Examples
1–1	Target idiom has identical or near-identical wording (lexis and/or syntactical structure) and meaning	‘bad apple’ → çürük elma (‘rotten apple’, Turkish); ცუდი ნაყოფი ‘cudi naqop’i’ (‘bad fruit’, Georgian); תפוח רקוב (‘rotten apple’, Hebrew); maçã podre (‘apple rotten’, Brazilian Portuguese). ‘add fuel to the fire’ → yangına körükle gitmek (‘to go to the fire with a bellows’, Turkish); hælde benzin på bå (‘pour gasoline to the fire’, Danish); ρίχνω λάδι στη φωτιά (‘throw oil on the fire’, Greek); deitar achas na fogueira (‘lay down woodsticks on the bonfire’, Portuguese); colocar lenha na fogueira (‘put woodsticks in the bonfire’, Brazilian Portuguese); įpilti aliejaus į ugnį (‘pour oil into the fire’, Lithuanian); afegir llenya al foc (‘add firewood to the fire’, Catalan). ‘apples and oranges’ → “epler og pærer” (‘apples and pears’, Norwegian); Äppel a Bieren (‘apples and pears’, Luxembourgish); jabolka in hruške (‘apples and pears’, Slovenian).‘big fish’ → დიდი მოთამაშე ‘didi motamashe’ (‘big player’, Georgian).
	Target idiom has different wording (lexis and/or syntactical structure) but same conceptual meaning	‘bad apple’ → (‘rotten goods’, Bulgarian). ‘eager beaver’ → /Allahtan bir gün oğurlamış (‘stealing one day from God’, Azeri). ‘ancient history’ → / köhnə palan (‘old packsaddle’, Azeri); acqua passata (‘water passed’, Italian). ‘beer and skittles’ → /təze küzə sərin su (‘new clay jug, cold water’, Azeri). ‘add fuel to the fire’ → /atash biyar-e mareke shodan (‘fire-bringer of the battlefield’, Farsi).
1–0	Target language uses descriptive (non-idiomatic) expression or has no equivalent	‘chocolate teapot’ → აბსოლუტურად უსარგებლო რამ (‘absolutely useless thing’, Georgian). ‘brain drain’ → /farar-e maghzhā (‘escape of brains’, Farsi). ‘agony aunt’ → conselheiro sentimental (‘sentimental counsellor’, Portuguese); მკითხველის მრჩეველი (‘reader’s advisor’, Georgian). ‘best man’ → apoio do noivo (‘support of the groom’, Portuguese); padrinho de casamento (‘wedding godfather’, Portuguese).
	Target language uses idiom made up of single lexical item (not an MWE)	‘brain teaser’ → pazli (Georgian); nøtt (‘nut’, Norwegian). ‘busy bee’ → futkari (Georgian). ‘brain teaser’ → /tappaca (Azeri); /moamma (Farsi). ‘cheat sheet’ → cola (Portuguese); skonaki (Greek); (‘pishtov’, Bulgarian); puska (Hungarian). ‘chicken feed’ → drobtinice (‘breadcrumbs’, Slovenian); shịshị (Igbo).
1–N	Source idiom has multiple Target idioms	‘brownie points’ → jó pont or pirospont (Hungarian). ‘brain drain’ → protų nutekėjimas; smegenų nutekėjimas (Lithuanian). ‘blind alley’ → /be Torkestan raftan (‘to go to Turkestan’, Farsi); /ab dar havan kubidan (‘pounding water in a mortar’, Farsi). ‘apples and oranges’ → /bir gazana atsan geynəməz (‘not to boil if placed together’, Azeri); /Allahları fərq eləməx (‘to have different Gods’, Azeri). ‘cloud nine’ → ברקיע השביעי, מאושר עד הגג (Hebrew).
N–1	Multiple source idioms have same Target idiom	‘big fish’ and ‘big cheese’ → peix gros (Catalan); didelė žuvis (Lithuanian); veľké zviera (Slovak); /böyük baş (Azeri). ‘big cheese’, ‘big wig’, ‘big fish’ → (‘vazhna klechka’, Bulgarian); stor kanon (Danish); /kalle gonde (Farsi); oke osisi (‘big tree’, Igbo). ‘big wig’ and ‘big cheese’ → liels čiekurs (Latvian); (Russian). ‘blind alley’ and ‘brick wall’ → מבוי סתום (Hebrew). ‘eager beaver’ and ‘busy bee’ → vreden kao pčela (Serbian); vreden kao pčela/mravka (Macedonian). ‘meat and potatoes’ and ‘bread and butter’ → arroz e feijão (Portuguese).

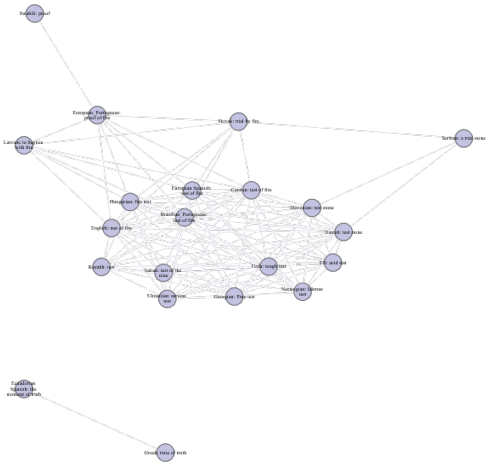
Table 2: Cross-linguistic mappings of English idioms showing different types of equivalence and adaptation.

concept as “mercato nero” in Italian (lit. “market black”). On the other hand, several expressions (e.g., “blue print”, “baby blues”, “box office”) often have no idiomatic equivalent, and instead their meaning is expressed through single words or non-idiomatic realisations.

Taken together, the two perspectives reveal that cross-linguistic idiomatic equivalence is not evenly distributed. Some PIEs are conceptually stable and widely share a lexicalisation across languages, while others are culture-specific or structurally divergent. For example, “beauty sleep” is near-literally shared across numerous languages (Figure 3a), whereas “acid test” exhibits a more fragmented pattern and tends to be rendered descriptively rather than idiomatically across languages (Figure 3b). Moreover, their meanings cannot be straightforwardly transferred from one language to the other simply by translating their individual words. Instead some language and culture specific knowledge is required, even among variants of the same language. This is the case of “bacia sem fundo” (lit. “basin without bottom”) in Brazilian Portuguese and “emplastro numa perna de pau” (lit. “plaster on a leg of wood”) in European Portuguese, both equivalent to “chocolate teapot” (someone or something which is useless) in English. This unevenness underscores the chal-



(a) *highly convergent* idiom (*beauty sleep*). A single large component with numerous connections among literal translations, indicating strong lexical transparency and widespread cross-linguistic lexicalisation.



(b) *divergent* example (*acid test*). One dominant component with smaller clusters, suggesting partial lexical convergence around a shared core metaphor.

Figure 3: Two representative idioms displaying distinct network structures.

lenges of modelling idiomaticity in multilingual NLP and highlights the need for resources that explicitly capture when and how idiomatic meanings are lexicalized differently across languages.

To capture patterns of cross-linguistic similarity among idiomatic expressions, we also represented each English PIE and its equivalents as graphs. For each English PIE, a graph was generated where each node corresponds to an equivalent language-specific PIE (represented in the graph via its literal translation to English), and where the edges represent lexical overlap between the PIEs in different languages. This means that two nodes are connected when, after stopword removal, they share at least one content word.

This heuristic can generate more than one subgraph for a given PIE, reflecting clusters of common lexicalisations. For any PIE, the number of nodes in the largest connected component of the graph reveals how many languages share, at least in part, lexicalisations. This means that a single-component graph would indicate that all languages share part of the PIE lexicalisation, while multiple sub graphs for a given PIE represent clusters of lexically related PIEs. Indeed, these graphs can highlight cross-lingual lexical overlaps between PIE equivalents in different languages, and Figure 3 shows the graphs for two representative PIEs. For example, “beauty sleep” is lexicalized using the multilingual equivalents of these words in a large variety of languages (Figure 3a). On the other hand, the resulting graphs may be disconnected, depending on how varied the lexicalisations of the idiomatic elements are across languages, as is the case for “acid test” (Figure 3b).

For each PIE, we computed structural graph measures that capture the overall connectivity and lexical cohesion across languages (see Table 3). These include the number of nodes and edges, the number of connected components, the proportion of nodes contained in the largest component, and the overall graph density. Edge weights quantify the amount of shared lexical material between translations, while the vocabulary size reflects the total number of distinct content words used. Finally, clique-based metrics identify subsets of languages whose literal translations are mutually related, providing a complementary view of complete lexical convergence.

Across the PIEs analysed, the graphs exhibit considerable structural variability. Most PIEs form fragmented networks (on average eight components per graph), indicating limited cross-linguistic lexical overlap. Only a few PIEs, such as “beauty sleep” and “black box”, display dense, single-component structures, reflecting high transparency and widespread lexicalisation.

The average density (0.31) confirms that idioms generally form partially connected clusters rather than a fully interconnected graph with a single lexicalisation across languages. The largest component typically contains about 60% of the nodes, showing that some PIEs maintain a common lexical core shared across many languages, while others are divided into several smaller, internally consistent groups. Clique metrics reinforce this distinction: most PIEs contain numerous small cliques, while only a few reach large clique numbers, corresponding to PIEs whose lexicalisation is shared across languages. Such groupings may reflect genealogical proximity or shared cultural norms, but the analysis remains agnostic as to their underlying causes. Two representative scenarios with distinct network structures are displayed in Figure 3.

Metric	Mean	SD	Min	Max	Interpretation
Nodes per PIE graph	12.0	8.9	4	53	Number of languages represented per idiom
Edges per PIE graph	37.2	77.4	1	18	Lexical connections between PIE equivalents
Number of components	5.0	3.1	1	18	Fragmentation of idiom networks
Largest component ratio	0.56	0.22	0.11	1.00	Proportion of nodes in the main cluster
Density	0.31	0.25	0.01	1.00	Overall connectivity among PIEs
Mean edge weight	1.55	0.46	1.0	3.33	Average number of shared words
Vocabulary size (unique words)	12.4	10.0	2	60	Lexical diversity in PIEs
Clique number	6.11	5.76	2	31	Size of the largest fully connected component
Number of maximal cliques	5.76	4.0	1	21	Number of fully connected subgraphs

Table 3: Summary statistics of graph-based measures across idioms.

5. Multilingual and Multimodal Idiomatic Representations

We analyze the XMPIE dataset and evaluate a publicly available vision–language baseline model, to showcase the challenges for language and vision models related to accurate idiomatic understanding. Among the challenges that this dataset can probe are those related to the multilingual idiomatic representation abilities of a model. In particular, the question of to what extent idiomatic representation is shared across languages, and if understanding in one language can lead to understanding in other languages. There are also questions related to multiple modalities, and to what degree accurate representation in one modality (like text) is also shared across modalities (like vision).

Lg	T1-I ↑	T1-L ↑	T2-I ↑	T2-L ↑	NDCG@5 ↑
EN	0.060	0.900	0.010	0.660	0.954
BP	0.100	0.580	0.020	0.360	0.910
ES	0.375	0.234	0.125	0.094	0.896
CN	0.140	0.316	0.018	0.140	0.877
TR	0.143	0.232	0.000	0.018	0.876

Table 4: PIE-only, no-training results with **EVA-CLIP-18B**. Top-2 (T2) is strict (both ranks must match). NDCG@5 uses gains $(1, 0.5, 0.5, 1, 0)$.

5.1. Experimental Setup

We evaluate systems on how well they rank five candidate images for each PIE. Each item provides five standardized image slots: one **idiomatic** image (image 1), one **literal** image (image 4), two **weak variants** (images 2–3), and one **distractor** (image 5).

This evaluation helps to determine if models are able to identify either the literal or the idiomatic target senses among the five images. This is an initial evaluation without sentences to provide contextual clues that could help to disambiguate between idiomatic and literal senses. Scores are computed per item and averaged per language.

5.2. Evaluation Metrics

Top-1 Accuracy (T1): We report two Top-1 settings: (i) **Idiomatic Top-1** (T1-I, correct if rank 1 is the idiomatic image); (ii) **Literal Top-1** (T1-L, correct if rank 1 is the literal image).

Top-2 Accuracy (T2): This metric is more strict and requires that both rank 1 and rank 2 correspond to the idiomatic and literal senses (or vice-versa).

Normalised Discounted Cumulative Gain (NDCG@5): To reflect graded usefulness across the whole list, we use NDCG@5. We assign gains based on the image slot:

$$g_{id} = (1,\ 0.5,\ 0,\ 0,\ 0), \quad g_{lit} = (0,\ 0,\ 0.5,\ 1,\ 0) \qquad \text{for slots } (1 \ldots 5).$$

For idiomatic instances, the idiomatic (image 1) is the most highly-weighted, with the part idiomatic (image 2) image receives partial credit (0.5). The unrelated (literal and distractor) images receive no credit (0). In literal instances, the same principle applies.

Given a system ranking π, where $\pi(i)$ denotes the image slot placed at rank i, the relevance at position i is defined as $rel_i = g_{\pi(i)}$. Pickard et al. (2025) report DCG; here we use the *normalised* variant to enable comparability across items. Let rel_i be the gain at rank i and $rel_i^{(\text{ideal})}$ the gain at rank i in the ideal ordering. Then

$$\text{NDCG@}5 \;=\; \frac{\sum_{i=1}^{5} \dfrac{rel_i}{\log_2(i+1)}}{\sum_{i=1}^{5} \dfrac{rel_i^{(\text{ideal})}}{\log_2(i+1)}} \in [0,1].$$

5.3. Baseline Method

We adopt **EVA-CLIP-18B** (Sun et al., 2024) as a no training, retrieval-style baseline, building on CLIP (Radford et al., 2021) because its joint image-text space enables simple, reproducible cosine-similarity ranking, and the large-scale EVA-CLIP-18B variant offers excellent retrieval performance.

For each item (PIE c), the PIE itself is used as the only text query and is compared against five candidate images in the model's joint embedding space.

Query Formulation: The text input is simply the PIE string c, with no prompt templating or additional context (e.g., "beauty sleep"). A single text embedding is computed for c and scored against the five images of the item.

Scoring and Ranking: Let $t = f_{\text{text}}(c)$ and $i_j = f_{\text{img}}(I_j)$ be EVA-CLIP-18B text and image embeddings. We L2-normalize (denoted by $\hat{i}_j$ and $\hat{t}$) and compute cosine scores

$$s_j = \hat{i}_j^\top \hat{t},$$

then rank images by s_j (descending); the top-ranked image is the prediction.

5.4. Results

We report PIE-only results for five languages with off-the-shelf pre-trained models without additional training or fine-tuning: English (EN), Brazilian Portuguese (BP), Ecuadorian Spanish (ES), Chinese (CN), and Turkish (TR). These languages were selected to represent a diverse language sample. For these we show *Idiomatic/Literal* Top-1, *Idiomatic/Literal* Top-2 accuracy, and the average NDCG@5 with symmetric gains $(1, 0.5, 0.5, 1, 0)$. NDCG@5 provides a single interpretable summary of ranking quality, while Top-1 preserves an intuitive success rate.

As shown in Table 4, **NDCG@5** is consistently high across languages (EN 0.954, BP 0.910, ES 0.896, CN 0.877, TR 0.876), indicating that—even without context—the model typically ranks the target and weak variants near the top. However, **Top-1** and especially the stricter **Top-2** reveal a clear targeting asymmetry: for EN/BP/CN/TR, the models consistently rank the *literal* image consistently, substantially outperforming the ranking of the *idiomatic* image (e.g., EN Top-1: 0.900 vs. 0.060; Top-2: 0.660 vs. 0.010). However, ES is the notable exception (idiomatic $>$ literal both on Top-1 and Top-2). The low idiomatic Top-2 values show that getting *both* the idiomatic image and its semantically-related distractor into the top two *in the correct order* is much harder than simply placing relevant images near the top (as NDCG suggests).

Examples: We observe several NDCG-perfect cases under the weak-half gains: EN—"pipe dream", "ghost town", "watering hole", "flying saucer"; BP—"colocar a boca no trombone", "mercado de pulgas"; ES—"vacaciones sin descanso", "sinverguenza", "premio academico"; CN—"黑箱"; TR—"büyük resim". Despite these strong cases, strict Top-2 remains low for idiomatic senses in most languages, confirming that fine-grained idiomatic disambiguation is still a bottleneck in model performance. Although in this evaluation only the PIE is provided, without any contextual clues for aiding the interpretation, the more relaxed evaluation measures adopted take this into account allowing any order of the target senses in the first positions as acceptable.

6. Conclusion

We presented XMPIE, a multilingual and multimodal idiomaticity dataset covering around 10K items for 34 language variants. This parallel resource allows cross-lingual analyses about the realisation of different concepts idiomatically providing insights into the salient linguistic and cultural aspects while also enabling assessment of the multilingual abilities of models and the extent to which understanding of an idiomatic expression transfers across languages and modalities.The results obtained with a baseline model on a subset of the dataset, reveal substantial cross-lingual variation and a consistent literal-over-idiomatic advantage in PIE-only experiments, underscoring the need for contextual cues for robust idiom understanding. Future work includes extending XMPIE and adding information about factors that may play a role in model and human processing like abstractness and imageability.

Data Release and Sharing Policy

In the era of LLMs, preventing evaluation benchmarks against data contamination during pretraining has become a critical challenge. A subset of the XMPIE dataset was recently used in the AdMIRe 2.0 shared task (Arslan et al., 2026), and the associated CodaBench competition will remain active as an ongoing benchmark to support systematic model evaluation. Additional language variants will be progressively integrated as their validation phases are completed.

To mitigate the risk of data leakage into future LLM training corpora, we adopt a phased and controlled data release strategy. For an initial period of one year, full access to the dataset will be restricted exclusively to the project contributors. Following this period, the dataset will be made available to the broader research community upon request. Access will require a brief proposal describing the intended use and agreement to terms of use that explicitly prohibit incorporating the dataset into public LLM training corpora.

Acknowledgments

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology). This work was also partly supported by UKRI (grants MR/U506734/1, EP/T02450X/1, and EP/S023062/1), National Council for Scientific and Technological Development (CNPq 406926/2025-5, 404722/2024- 5; 313103/2021-6), Minas Gerais State Agency for Research and Development (FAPEMIG) and EQUATE, CA23147 COST action GOBLIN, industry partners and the Research Council of Norway with funding to MediaFutures: Research Centre for Responsible Media Technology and Innovation, through the centers for Research-based Innovation scheme (project number 309339).

Ethical Considerations and Limitations

Item Vetting: During content creation, annotators could (i) mark items as successfully generated; (ii) flag problematic items with notes; or (iii) propose removal. Removal reasons included: always-literal expressions, obsolete items, offensive content, or cases where a viable literal rendering was unattainable.

Safety Filters: Guidelines explicitly avoided swearing, illegal activity, and negatively framed mentions of specific people/organisations in context sentences; image prompts similarly steered clear of upsetting or harmful content.

Cross-lingual Effects: Because for nuanced language image generation models seem to perform best with English prompts, contributors translated non-English prompts where helpful, while preserving language-specific idiomatic content in the contexts and item selection. Therefore there may be nuanced effects that are not represented given the cultural language bias in the image generation models.

Language Validation and Coverage: As detailed in Section 4, all 34 language variants underwent rigorous human review. However, they may differ in their coverage of potentially idiomatic expressions and in the completeness of the different modalities. Cross-lingual analyses and performance metrics involving languages in the Extended Dataset should be interpreted with this variance in mind, as these subsets may lack certain elements (such as context sentences) or exhibit heterogenous qualitative consistency compared to the Core Dataset.

7. Bibliographical References

Sophia Ananiadou, John McNaught, and Paul Thompson. 2012. *The English Language in the Digital Age*. META-NET White Paper Series. Springer Nature, Heidelberg. EC FP7 PSP Project METANET4U.

Doğukan Arslan, Rodrigo Wilkens, Wei He, Dilara Torunoğlu-Selamet, Thomas Pickard, Aline Villavicencio, Adriana S. Pagano, and Gülşen Eryiğit. 2026. MWE-2026 Shared Task 2: AdMIRe 2 - Advancing Multimodal Idiomaticity Representation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morocco. Association for Computational Linguistics.

Joanne Boisson, Luis Espinosa-Anke, and Jose Camacho-Collados. 2023. Construction artifacts in metaphor identification datasets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 6581–6590, Singapore. Association for Computational Linguistics. Main Conference.

Jonathan Charteris-Black. 2002. Second language figurative proficiency: A comparative study of Malay and English. *Applied Linguistics*, 23(1):104–133.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics. ACL 2022.

Wei He, Tiago Kramer Vieira, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2025. Investigating idiomaticity in word representations. *Computational Linguistics*, 51(2):505–555.

Suzanne Irujo. 1986. Don't put your leg in your mouth: Transfer in the acquisition of idioms in a second language. *TESOL Quarterly*, 20(2):287–304.

Zoltan Kovecses. 2010. *Metaphor: A practical introduction*. Oxford university press.

Christopher Moseley. 2010. *Atlas of the world's languages in danger*, 3 edition. UNESCO Publishing, Paris.

Dylan Phelps, Thomas Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE–UD 2024) at LREC–COLING 2024*, pages 178–187, Turin, Italy. ELRA and ICCL. Workshop paper.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Agata Savary, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine De Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, et al. 2024. Unidive: A cost action on universality, diversity and idiosyncrasy in language technology. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Underresourced Languages@ LREC-COLING 2024*, pages 372–382.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resources Association (ELRA).

Arkaitz Zubiaga. 2024. Natural language processing in the era of large language models: A survey. *Frontiers in Artificial Intelligence*, 6:1350306.

8. Language Resource References

Doğukan Arslan, Hüseyin Anıl Çakmak, Gülşen Eryiğit, and Joakim Nivre. 2025. Using LLMs to advance idiom corpus construction. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 21–31, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-tokens dataset. *Towards a Shared Task for Multiword Expressions (MWE 2008)*, page 19.

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

Gülşen Eryiğit, Ali Şentaş, and Johanna Monti. 2022. Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering*, 29(4):909–941.

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.

Chikara Hashimoto and Daisuke Kawahara. 2009. Compilation of an idiom example database for supervised idiom identification. *Language Resources and Evaluation*, 43(4):355–384.

Wei He, Tiago Kramer Vieira, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2025. Investigating idiomaticity in word representations. *Computational Linguistics*, 51(2):505–555.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, pages 10258–10279, Miami, Florida, USA. Association for Computational Linguistics. Main Conference.

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.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2025. Rolling the DICE on idiomaticity: How LLMs fail to grasp context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7314–7332, Vienna, Austria. Association for Computational Linguistics.

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, and Marco Idiart. 2025. Semeval-2025 task 1: Admire – advancing multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2597–2609, Vienna, Austria. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

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.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesús-Germán Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Naome Etori, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, volume 37, pages 11479–11505. Curran Associates, Inc. Datasets and Benchmarks Track.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2025. Understanding figurative meaning through explainable visual entailment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1–23, Albuquerque, New Mexico. Association for Computational Linguistics.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Mathieu Constant, Petya Osenova, and Federico Sangati. 2015. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland.

Prateek Saxena and Soma Paul. 2020. Epie dataset: A corpus for possible idiomatic expressions. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings*, page 87–94, Berlin, Heidelberg. Springer-Verlag.

Uliana Sentsova, Debora Ciminari, Josef Van Genabith, and Cristina España-Bonet. 2025. MultiCoPIE: A multilingual corpus of potentially idiomatic expressions for cross-lingual PIE disambiguation. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 67–81, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of*

the ACL (EACL 2009), pages 754–762, Athens, Greece. Association for Computational Linguistics.

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).

Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. 2024. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022a. 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, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022b. 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 pretrained 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*, page 2715–2726. Association for Computational Linguistics.

Xiaoyu Tong, Zhi Zhang, Martha Lewis, and Ekaterina Shutova. 2025. Hummus: A dataset of humorous multimodal metaphor use.

Tongguan Wang, Mingmin Wu, Guixin Su, Dongyu Su, Yuxue Hu, Zhongqiang Huang, and Ying Sha. 2025. MChIRC: A multimodal benchmark for Chinese idiom reading comprehension. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. IRFL: Image recognition of figurative language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1044–1058, Singapore. Association for Computational Linguistics.

Appendix A. Affiliation List

Table 5 lists the authors and their corresponding affiliations.

Author	Affiliation	Author	Affiliation
Dilara Torunoğlu-Selamet	Istanbul Techical University	Doğukan Arslan	Istanbul Techical University
Rodrigo Wilkens	University of Exeter	Wei He	University of Exeter
Doruk Eryiğit	ITU NLP	Thomas Pickard	University of Sheffield
Adriana S. Pagano	Federal University of Minas Gerais	Aline Villavicencio	University of Exeter
Gülşen Eryiğit	Istanbul Techical University	Ágnes Abuczki	Károli Gáspár University of the Reformed Church in Hungary
Aida Cardoso	Research Center for Linguistics at NOVA University Lisbon	Alesia Lazarenka	Tesi s.r.l.
Dina Almassova	Nazarbayev University	Amalia Mendes	University of Lisbon
Anna Kanellopoulou	Aristotle University of Thessaloniki	Antoni Brosa-Rodríguez	Universitat Rovira i Virgili
Baiba Valkovska	Institute of Mathematics and Computer Science, University of Latvia	Beata Wojtowicz	University of Warsaw
Bolette Pedersen	University of Copenhagen	Carlos Manuel Hidalgo-Ternero	University of Malaga, IUITLM
Chaya Liebeskind	Jerusalem College of Technology	Danka Jokić	University of Belgrade
Diego Alves	Saarland University	Eleni Triantafyllidi	Aristotle University of Thessaloniki
Erik Velldal	University of Oslo	Fred Philippy	University of Luxembourg
Giedre Valunaite Oleskeviciene	Mykolas Romeris University	Ieva Rizgeliene	Institute of Data Science and Digital Technologies, Vilnius University
Inguna Skadina	Institute of Mathematics and Computer Science, University of Latvia	Irina Lobzhanidze	Ilia State University
Isabell Stinessen Haugen	University of Bergen	Jauza Akbar Krito	Universitas Gadjah Mada
Jelena M. Marković	University of East Sarajevo	Johanna Monti	University of Naples L'Orientale
Josue Alejandro Sauca	International University of Valencia	Kaja Dobrovoljc	University of Ljubljana & Jozef Stefan Institute
Kingsley O. Ugwuanyi	SOAS University of London	Laura Rituma	Institute of Mathematics and Computer Science, University of Latvia
Lilja Øvrelid	University of Oslo	Maha Tufail Agro	Mohamed bin Zayed University of Artificial Intelligence

Continued on next page

Table 5 – Continued from previous page

Author	Affiliation	Author	Affiliation
Manzura Abjalova	Alisher Navoi' Tashkent State University of Uzbek Language and Literature	Maria Chatzigrigoriou	National and Kapodistrian University of Athens
María del Mar Sánchez Ramos	University of Alcalá	Marija Pendevska	St. Cyrillus and Methodius University
Masoumeh Seyyedrezaei	Istinye University	Mehrnoush Shamsfard	Shahid Beheshti University
Momina Ahsan	MBZUAI	Muhammad Ahsan Riaz Khan	MBZUAI
Nathalie Carmen Hau Norman	University of Copenhagen	Nilay Erdem Ayyıldız	Fırat University
Nina Hosseini-Kivanani	University of Luxembourg & RTL	Noémi Ligeti-Nagy	ELTE Research Centre for Linguistics
Numaan Naeem	MBZUAI	Olha Kanishcheva	Heidelberg University & SET University
Olha Yatsyshyna	Ternopil Volodymyr Hnatiuk National Pedagogical University	Daniil Orel	MBZUAI
Petra Giommarelli	University of Pisa & University of Naples L'Orientale	Petya Osenova	Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
Radovan Garabik	Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences	Regina E. Semou	National and Kapodistrian University of Athens
Rozane Rebechi	Federal University of Rio Grande do Sul	Salsabila Zahirah Pranida	MBZUAI
Samia Touileb	University of Bergen	Sanni Nimb	Society for Danish Language and Literature
Sarfraz Ahmad	MBZUAI	Sarvinoz Sharipova	Samarkand State Institute of Foreign Languages
Shahar Golan	Jerusalem College of Technology	Shaoxiong Ji	ELLIS Institute Finland and University of Turku
Sopuruchi Christian Aboh	The Hong Kong Polytechnic University	Srdjan Sucur	University of East Sarajevo
Stella Markantonatou	ILSP and Archimedes, Athena RC	Sussi Olsen	University of Copenhagen
Vahide Tajalli	NLP Research Lab, Shahid Beheshti University	Veronika Lipp	ELTE Research Centre for Linguistics
Voula Giouli	Aristotle University of Thessaloniki	Yelda Yeşildal Eraydın	Fırat University
Zahra Saaberi	NLP Research Lab, Shahid Beheshti University	Zhuohan Xie	MBZUAI

Table 5: Full list of authors and their corresponding institutions.