A Survey of Idiom Datasets for Psycholinguistic and Computational Research

Michael Flor
Educational Testing Service
Princeton, New Jersey, USA
mflor@ets.org

Xinyi Liu
Montclair State University
Montclair, New Jersey, USA
liux2@montclair.edu

Anna Feldman
Montclair State University, Montclair, New Jersey, USA
feldmana@montclair.edu

Abstract

Idioms are figurative expressions whose meanings often cannot be inferred from their individual words, making them difficult to process computationally and posing challenges for human experimental studies. This survey reviews datasets developed in psycholinguistics and computational linguistics for studying idioms, focusing on their content, form, and intended use. Psycholinguistic resources typically contain normed ratings along dimensions such as familiarity, transparency, and compositionality, while computational datasets support tasks like idiomaticity detection/classification, paraphrasing, and cross-lingual modeling. We present trends in annotation practices, coverage, and task framing across 53 datasets. Although recent efforts expanded language coverage and task diversity, there seems to be no relation yet between psycholinguistic and computational research on idioms.

1 Introduction

Idioms are conventional expressions whose meanings cannot be reliably inferred from the meanings of their individual words. Phrases such as *kick the bucket* or *spill the beans* demonstrate how idioms often convey figurative meanings that diverge from literal interpretations. Because of their semantic opacity and syntactic variability, idioms present persistent challenges for natural language processing (NLP) systems.

Recognizing and interpreting idiomatic expressions is essential for a range of NLP applications, including machine translation, dialogue systems, and sentiment analysis. However, automatic idiom processing remains difficult. The same expression can be used literally or figuratively depending on context, and idioms vary in their degree of flexibility, compositionality, and transparency. Addressing these challenges requires high-quality datasets that capture the complexities of idiomatic language in realistic contexts.

Over the past two decades, a number of datasets for idiom processing have been created, often as part of individual research projects. However, these resources are highly heterogeneous: they differ in their annotation schemes, languages covered, idiom types included, and dataset sizes. There is currently no standard benchmark, no unified annotation framework, and limited cross-dataset compatibility, making it difficult to compare methods or advance the field systematically.

This paper surveys the available datasets for idiom analysis and processing. Those come from two different disciplines - psycholinguistics and computational linguistics. We provide an overview of their properties, including language coverage, annotation strategies, intended tasks, and dataset sizes. We highlight the strengths and limitations of existing resources and identify areas where further development is needed to support progress in idiom-aware NLP. A continuously updated list of idiom datasets, with links to datasets and publications, is maintained at https://github.com/maafiah/IdiomsResearch.

2 Idiom datasets in psycholinguistic research

In psycholinguistics, research on idioms has been largely focused on the immediate processing of idioms, such as reading time, most notably the comprehension of idioms in (and out of) context. The motivations for specific hypotheses in such research are often related to the linguistically-informed theories or models of idiom processing (Cacciari, 2014; Espinal and Mateu, 2019).

The time-course of idiom processing by human subjects is influenced by a variety of variables, such as idiom familiarity. Controlling for such variables is an important feature of psycholinguistic experiments. Some of those variables are motivated by linguistic theories, other variables are motivated

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 90–100, 2025

by psychological considerations. Norming studies allow researchers to collect subjective individual ratings for a variety of postulated dimensions and aggregate them across respondents (Winter, 2022). Idiom norming studies use Likert scales to rate idiomatic expressions on various dimensions. However, the constructs are not always the same across studies. Table 1 in the appendix lists 16 public datasets dedicated to psycholinguistic aspects of idiomatic expressions. Most of those datasets are publicly available. The aspects/dimensions are described below.

Familiarity of an idiom has been construed as the frequency with which a person has been exposed to a given expression in their everyday life. However, since such frequency cannot be measured objectively, the standard approach is to ask raters for a subjective estimation (*subjective frequency*), explained as 'familiarity'. Several studies demonstrated that familiar idioms are processed faster than less familiar ones (Schweigert, 1986; Schweigert and Moates, 1988).

Knowledge of meaning. A notion related to familiarity is knowledge, the degree to which a person thinks they know the meaning of an idiom, and can explain what the expression means (Li et al., 2016). This also sometimes serves as a control on the adequacy of participant ratings (Pagliai, 2023).

The notion of **literality** (also called literalness or ambiguity) concerns whether the idiom has a plausible literal interpretation. For example, *break the ice* has a literal interpretation, but *shoot the breeze* is semantically anomalous because a breeze is not something that can be shot. Literality is usually measured by asking participants to rate whether the phrase could be used literally in addition to its figurative meaning (Libben and Titone, 2008; Caillies, 2009; Tabossi et al., 2011). While literality could be conceived as a dichotomous variable, it is usually measured on a scale.

Compositionality and **decomposability**. This dimension estimates the degree to which the idiom's component words contribute to its idiomatic interpretation. This is based on the notion that some idioms can be compositionally analyzed (Nunberg et al., 1994). For example, for *spill the beans*, meaning 'reveal secrets', the verb 'spill' corresponds to 'reveal', 'beans' corresponds to 'secrets', showing some composition of the whole figurative meaning. No such composition occurs in e.g. *kick the bucket*, that idiom is not decomposable. Hamblin and Gibbs (1999) suggested that the degree to which the meaning of idiom parts contributes to idiom interpretation can be classified on a continuum, calling it *degree of analyzability*. Experimental studies have found mixed effects of idiom analyzability on processing time (Titone and Libben, 2014; Tabossi et al., 2008).

Transparency is defined as the ease by which the motivation for the structure of an idiomatic expression can be deduced (Nunberg et al., 1994), or how easy it is to recognize why an idiom means what it means. Such motivation can be a metaphorical relation (e.g., the idiom *flip one's lid* might be based on a metaphor that anger is like steam), or a historical remnant (e.g., '*carry coals to Newcastle*'), etc. The reciprocal notion for transparency is *opacity*. Semantic (or conceptual) transparency of idioms is not the same notion as compositionality, but they are related as the ratings are often highly correlated. According to Citron et al. (2016), transparency is a problematic measure since participant ratings are based on intuitions and guesses, and highly dependent on the knowledge of the correct idiomatic meaning.

Age of acquisition (AoA) is the estimated age at which a word (or expression) and its meaning were first learned (Johnston and Barry, 2006). AoA is typically measured in years and months. While AoA for words has been a common measure in psycholinguistics (Morrison et al., 1997), norming AoA estimations for idiomatic expressions was introduced by Tabossi et al. (2011). AoA was found to correlate with knowledge, familiarity, and subjective frequency for idioms (Bonin et al., 2013, 2018; Li et al., 2016). Li et al. (2016) suggested that the earlier an idiom is learned, the more frequently it might be encountered, and thus become more familiar. Bonin et al. (2013) found that AoA was predictive for idiom reading times. Bonin et al. (2018) suggested that AoA norms might be useful for studying idiom processing in children.

Predictability of idioms is defined as the probability of completing an incomplete string to a full idiomatic expression, for example completing *'be in seventh....'* with *'heaven'*. For measuring predictability, participants are asked to read incomplete sentences with idioms and provide the last word of the idiom (that is blanked out).

Syntactic flexibility refers to the notion that idiomatic expressions are often not entirely frozen and allow some degrees of syntactic variability,

such as inflection of verbs, insertions of adjective or adverbs, sometimes passivization, etc., (Fraser, 1970; Moon, 1998). As noted by Gibbs and Gonzales (1985), this can have psychological implications. For example, flexible idioms can easily undergo transformations, e.g., for *throw in the towel*, the passive form *the towel was thrown in by him* retains the idiomatic meaning *he gave up*. Less flexible expressions are less likely to be interpreted figuratively, e.g., *the bucket was kicked by him* is less likely to be interpreted as *he died*. For norming studies, idiomatic expressions are presented in various forms and participants are asked to rate their idiomaticity (Tabossi et al., 2011).

The notion of **concreteness** describes the extent to which the meaning of a word refers to a state or event that can be experienced via one or more sensory modalities (Paivio et al., 1968). In experiments, participants process concrete words faster and more accurately than abstract words (the so called *concreteness effect*, Paivio 1991). Two studies collected concreteness ratings for idiomatic phrases, as opposed to concreteness of constituent words (Citron et al., 2016; Morid and Sabourin, 2024). Whether concreteness of idioms might be related to idiom processing or representation is yet unknown.

Imageability refers to the ability to create a mental image of a word. Generally, imageability enhances word recognition (Connell and Lynott, 2012). Imageability is often highly correlated with concreteness. Mental imagery has been linked to idiom comprehension (Gibbs and O'Brien, 1990), and norming imageability of idioms is a new research trend.

In psycholinguistics, three important non-cognitive aspects are known to have influence on representation and meaning of individual words (Osgood et al., 1957; Russell, 2003), known collectively as VAD: **valence** (affective value, sentiment, positiveness–negativeness, or pleasantness of a stimulus), **arousal** (feeling active or passive, or intensity associated with the stimulus), and **dominance** (dominant–submissive, degree of control). VAD norms for individual words are well known in psycholinguistic research (e.g. Warriner et al. 2013), and also in computational linguistics research (Mohammad, 2018). Nunberg et al. (1994) stated that "idioms are typically used to imply a certain evaluation or affective stance toward the things they denote". Peng et al. (2014) used the rated valence of idiom component words for automated detection of idioms in texts. Obtaining valence and arousal ratings of whole idiomatic expressions is a recent trend in norming studies (Gavilán et al., 2021; Morid and Sabourin, 2024).

3 Idiom datasets in computational linguistic research

Computational linguistics research on idioms has produced a wide range of datasets designed for classification, disambiguation, paraphrasing, and multilingual modeling. Unlike psycholinguistic norming studies, which focus on controlled ratings of idiom properties, these resources are typically drawn from real-world corpora and are suited for NLP tasks. Table 2 in the appendix lists the datasets published in computational-linguistics literature. Most of those datasets are publicly available.

3.1 Classification and Disambiguation

A significant portion of computational idiom datasets focus on binary classification – distinguishing idiomatic from literal uses of the same expression in context. The VNC-Tokens dataset (Fazly et al., 2009) includes nearly 3,000 usages of verb-noun combinations in English, annotated as idiomatic or literal. Similarly, the IDIX corpus (Sporleder et al., 2010) collected over 5,800 English sentences labeled for idiomaticity, though it is not publicly available.

Several benchmark datasets have been created in the context of shared tasks. SemEval-2013 Task 5b (Korkontzelos et al., 2013) presents 85 ambiguous idioms with over 4,000 instances, each provided with a five-sentence context for disambiguation. The RU Idioms dataset (Aharodnik et al., 2018) includes 2,420 idiomatic instances and 3,027 literal ones, drawn from Classical and Modern Russian literature and Wikipedia texts. Designed to support supervised classification of idiomaticity in Russian, the corpus provides richly contextualized examples in paragraph-level texts. The MAGPIE dataset (Haagsma et al., 2020) introduced 56,622 instances of potential idioms in short textual context, for 1,756 different English idioms.

3.2 Paraphrase, Sentiment, and Substitution

Beyond disambiguation, idiom paraphrasing has been a central task. Pershina et al. (2015) collected over 2,400 English idioms with associated paraphrases and 1,400 idiom-idiom pairs annotated

for mutual paraphrasability. The Idiom Substitution dataset (Liu and Hwa, 2016) offers definitions and plausible substitutions for 172 English idioms, useful for generation tasks. The Parallel Idioms Corpus (Zhou et al., 2021) presented 823 English-language idioms with 5,170 sentence-pairs, where a sentence contains an idiom or its literal paraphrase.

Sentiment-oriented datasets such as IDIOMENT (Williams et al., 2015) and SLIDE (Jochim et al., 2018) capture affective dimensions. IDIOMENT includes 580 idioms with both idiomatic and sentence-level sentiment annotations, while SLIDE contains sentiment scores for over 5,000 idioms. IDEM (Prochnow et al., 2024) has about 9685 sentences with idioms and labels for expressed emotion in each sentence. These resources are especially relevant for tasks like figurative sentiment classification and sentiment-aware generation.

3.3 Multilingual Resources

Recent efforts have addressed the lack of multilingual idiom data. The LIdioms dataset (Moussallem et al., 2018) provides 815 idioms across five languages (English, German, Italian, Portuguese, and Russian) in RDF format, linking semantically equivalent expressions. The IMIL corpus (Agrawal et al., 2018) contains over 2,200 English idioms and their translations in seven Indian languages, annotated across 250K sentences. PETCI (Tang, 2022) includes 4,310 Chinese idioms with 29,936 English translations, capturing diverse translation errors and paraphrase strategies.

SemEval-2022 Task 2 (Tayyar Madabushi et al., 2022) extended idiomaticity detection to three languages – English, Portuguese, and Galician – by providing labeled training and development data for English and Portuguese, and zero-shot test data for Galician. While this marked a step toward multilingual idiom processing, the dataset does not include aligned idiom instances across languages or annotations for contextual factors such as register, familiarity, or cultural variation, limiting its utility for studying pragmatic differences.

IdiomKB (Li et al., 2024b) merged several previously published idiom datasets (for English, Chinese, and Japanese), with the purpose of improving cross-lingual LLM-based translation of texts with idiomatic expressions.

3.4 Model Probing and Representation Learning

Several datasets have been developed to probe and improve language models' handling of idioms. AStitchInLanguageModels (Tayyar Madabushi et al., 2021) comprises naturally occurring sentences containing potentially idiomatic multiword expressions (MWEs) in English and Portuguese, annotated with fine-grained meanings and paraphrases. This dataset supports tasks evaluating models' ability to detect idiom usage and generate effective representations of idiomaticity.

IDIOMEM (Haviv et al., 2023) is a probe dataset of English idioms used to analyze memorization behavior in transformer language models. It facilitates the study of when and how models recall memorized idiomatic sequences. CultureLLM (Li et al., 2024a) focuses on incorporating cultural differences into large language models (LLMs), generating semantically equivalent training data through semantic data augmentation, fine-tuning culture-specific LLMs for nine cultures. Liu et al. (2024) and Khoshtab et al. (2024) use proverbs and idioms to probe LLMs inference in processing figurative language across multiple languages.

3.5 Gaps and Future Directions

While these datasets have enabled significant progress in idiom-aware NLP, several limitations persist. Annotation schemes vary widely across datasets—some mark idiomaticity at the phrase level, others at the sentence level; some provide paraphrases or sentiment labels, others do not. This heterogeneity reflects the diversity of research goals, but it also makes cross-dataset evaluation difficult. Dataset sizes and scopes differ considerably: some focus on depth (many instances per idiom), others on breadth (many idioms with few examples). While multilingual coverage is growing, most datasets do not include semantically aligned idioms across languages or cultural contexts. In addition, idioms are drawn from limited domains, with little attention to genre, register, or discourse variation. Rather than imposing a single standard, future work could benefit from shared metadata conventions, interoperable formats, and more transparent documentation to support comparison, reuse, and integration across datasets. In addition, there seems to be no relation yet between psycholinguistic and computational research on idioms. This is not surprising, as those research domains tradition-

ally had different orientations and research agendas. Notably, datasets from psycholinguistics focus on idioms as 'types', while computational datasets often work with idioms as 'tokens' (instances) in various textual contexts. One area where some cross-pollination between the fields may develop is the notion of valence (sentiment) of idioms. Another possible direction could be using computational methods to model/explain human ratings on various aspects of idioms.

4 Conclusion

This survey compares idioms datasets developed in psycholinguistics and computational linguistics, focusing on their design, intended use, and underlying assumptions. By examining both types, we highlight how methodological differences shape what each dataset can support.

Limitations

While we conducted an extensive search for various research-oriented datasets of idiomatic expressions, some resources might have been missed.

Acknowledgments

This material is partially based upon work supported by the U.S. National Science Foundation under Grant Numbers 2428506 and 2226006.

References

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).

Sara D. Beck. 2020. *Native and Non-native Idiom Processing: Same Difference*. Ph.D. thesis, University of Tübingen.

Sara D. Beck and Andrea Weber. 2016. English-German Database of Idiom Norms. University of Tübingen.

Patrick Bonin, Alain Méot, Jean-Michel Boucheix, and Aurélia Bugaiska. 2018. Psycholinguistic norms for 320 fixed expressions (idioms and proverbs) in French. *Quarterly Journal of Experimental Psychology*, 71(5):1057–1069.

Patrick Bonin, Alain Méot, and Aurélia Bugaiska. 2013. Norms and comprehension times for 305 French idiomatic expressions. *Behavior Research Methods*, 45:1259–1271.

Nyssa Z. Bulkes and Darren Tanner. 2017. "Going to town": Large-scale norming and statistical analysis of 870 American English idioms. *Behavior Research Methods*, 49:772–783.

Cristina Cacciari. 2014. Processing multiword idiomatic strings: Many words in one? *The Mental Lexicon*, 9(2):267–293.

Stéphanie Caillies. 2009. Descriptions de 300 expressions idiomatiques: familiarité, connaissance de leur signification, plausibilité littérale, «décomposabilité» et «prédictibilité». *L'Année psychologique*, 109:463–508.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Francesca M. M. Citron, Cristina Cacciari, Michael Kucharski, Luna Beck, Markus Conrad, and Arthur M. Jacobs. 2016. When emotions are expressed figuratively: Psycholinguistic and Affective Norms of 619 Idioms for German (PANIG). *Behavior Research Methods*, 48:91–111.

Louise Connell and Dermot Lynott. 2012. Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition*, 125(3):452–465.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The vnc-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, Marrakech, Morocco.

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

Ricarda Dormeyer and Ingrid Fischer. 1998. Building Lexicons out of a Database for Idioms. In *Proceedings of the First International Conference on Language Resources and Evaluation*, pages 833–838.

M. Teresa Espinal and Jaume Mateu. 2019. Idioms and Phraseology. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

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).

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Christiane Fellbaum and Alexander Geyken. 2005. Transforming a Corpus Into a Lexical Resource - the Berlin Idiom Project. *Revue française de linguistique appliquée*, X(2):49–62.

Fraser. 1970. Idioms within a transformational grammar. *Foundation of Language*, 6:22–42.

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.

José M. Gavilán, Juan Haro, José Antonio Hinojosa, Isabel Fraga, and Pilar Ferré. 2021. Psycholinguistic and affective norms for 1,252 Spanish idiomatic expressions. *PLoS ONE*, 16(7).

Raymond W. Gibbs and Gayle P. Gonzales. 1985. Syntactic frozenness in processing and remembering idioms. *Cognition*, 20(3):243–259.

Raymond W. Gibbs and Jennifer E. O'Brien. 1990. Idioms and mental imagery: The metaphorical motivation for idiomatic meaning. *Cognition*, 36(1):35–68.

Begoña Góngora, Andre Gómez-Lombardi, and Alonso Ortega González. 2022. Descriptive Norms for 1,082 Chilean-Spanish Idiomatic Expressions. *Revista signos*, 55(110).

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.

Jennifer L. Hamblin and Raymond W. Gibbs. 1999. Why You Can't Kick the Bucket as You Slowly Die: Verbs in Idiom Comprehension. *Journal of Psycholinguistic Research*, 28:25–39.

Chikara Hashimoto and Daisuke Kawahara. 2008. Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 992–1001, Honolulu, Hawaii. Association for Computational Linguistics.

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.

Ferdy Hubers, Catia Cucchiarini, Helmer Strik, and Ton Dijkstra. 2019. Normative Data of Dutch Idiomatic Expressions: Subjective Judgments You Can Bank on. *Frontiers in Psychology*, 10:1075.

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 *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Robert A. Johnston and Christopher Barry. 2006. Age of acquisition and lexical processing. *Visual Cognition*, 13(7-8):789–845.

Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2024. Comparative study of multilingual idioms and similes in large language models. *Preprint*, arXiv:2410.16461.

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.

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.

Anastasia Lada, Philippe Paquier, Ifigenia Dosi, Christina Manouilidou, Simone Sprenger, and Stefanie Keulen. 2024. Four hundred Greek idiomatic expressions: Ratings for subjective frequency, ambiguity, and decomposability. *Behavior Research Methods*, 56:8181–8195.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. CultureLLM: Incorporating Cultural Differences into Large Language Models. *Preprint*, arXiv:2402.10946.

Degao Li, Yu Zhang, and Xiaolu Wang. 2016. Descriptive norms for 350 Chinese idioms with seven syntactic structures. *Behavior Research Methods*, 48:1678–1693.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024b. Translate meanings, not just words: IdiomKB's role in optimizing idiomatic translation with language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Maya R. Libben and Debra A. Titone. 2008. The multidetermined nature of idiom processing. *Memory & Cognition*, 36(6):1103–1123.

Changsheng Liu and Rebecca Hwa. 2016. Phrasal substitution of idiomatic expressions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 363–373, San Diego, California. Association for Computational Linguistics.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Rosamund Moon. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Clarendon Press, Oxford.

Mahsa Morid and Laura Sabourin. 2024. Affective and sensory–motor norms for idioms by L1 and L2 English speakers. *Applied Psycholinguistics*, 45(1):138–155.

Catriona M. Morrison, Tameron D. Chappell, and Andrew W. Ellis. 1997. Age of Acquisition Norms for a Large Set of Object Names and Their Relation to Adult Estimates and Other Variables. *The Quarterly Journal of Experimental Psychology Section A*, 50(3):528–559.

Diego Moussallem, Mohamed Ahmed Sherif, Diego Esteves, Marcos Zampieri, and Axel-Cyrille Ngonga Ngomo. 2018. LIdioms: A multilingual linked idioms data set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Emily Nordmann and Antonia A. Jambazova. 2017. Normative data for idiomatic expressions. *Behavior Research Methods*, 49:198–215.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. 1957. *The measurement of meaning*. University of Illinois Press.

Irene Pagliai. 2023. Bridging the Gap: Creation of a Lexicon of 150 Pairs of English and Italian Idioms Including Normed Variables for the Exploration of Idiomatic Ambiguity. *Journal of Open Humanities data*, 9(1).

Allan Paivio. 1991. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, 45(3):255–287.

Allan Paivio, John C. Yuille, and Stephen A. Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 76(1):1–25.

Jing Peng, Anna Feldman, and Ekaterina Vylomova. 2014. Classifying idiomatic and literal expressions using topic models and intensity of emotions. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2019–2027, Doha, Qatar. 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.

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.

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.

James A. Russell. 2003. Core affect and the psychological construction of emotio. *Psychological Review*, 110(1):145–172.

Prateek Saxena and Soma Paul. 2021. Labelled epie: A dataset for idiom sense disambiguation. In *Text, Speech, and Dialogue*, pages 210–221, Cham. Springer International Publishing.

Wendy A. Schweigert. 1986. The comprehension of familiar and less familiar idioms. *Journal of Psycholinguistic Research*, 15(1):33–45.

Wendy A. Schweigert and Danny R. Moates. 1988. Familiar idiom comprehension. *Journal of Psycholinguistic Research*, 17(4):281–296.

Marco Silvio Giuseppe Senaldi. 2019. *Working both sides of the street: computational and psycholinguistic investigations on idiomatic variability*. Ph.D. thesis, Scuola Normale Superiore, Pisa.

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).

Patrizia Tabossi, Lisa Arduino, and Rachele Fanari. 2011. Descriptive norms for 245 Italian idiomatic expressions. *Behavior Research Methods*, 43:110–123.

Patrizia Tabossi, Rachele Fanari, and Kinou Wolf. 2008. Processing idiomatic expressions: Effects of semantic compositionality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(2):313–327.

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.

Debra Titone and Maya Libben. 2014. Time-dependent effects of decomposability, familiarity and literal plausibility on idiom priming: A cross-modal priming investigation. *The Mental Lexicon*, 9(3):473–496.

Lei Wang and Shiwen Yu. 2010. Construction of Chinese idiom knowledge-base and its applications. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 11–18, Beijing, China. Coling 2010 Organizing Committee.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Lowri Williams, Christian Bannister, Michael Arribas-Ayllon, Alun Preece, and Irena Spasić. 2015. The role of idioms in sentiment analysis. *Expert Systems with Applications*, 42(21):7375–7385.

Bodo Winter. 2022. Managing Semantic Norms for Cognitive Linguistics, Corpus Linguistics, and Lexicon Studies. In Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister, editors, *The Open Handbook of Linguistic Data Management*, pages 489–497. MIT Press.

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 778–787, 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 Appendix

In this appendix we list datasets from psycholinguistic research in Table 1, and datasets from computational linguistics research in Table 2.

Authors	Count	Language	Contents (rating dimensions)	Avail.
Libben and Titone (2008)	210	English	familiarity, meaningfulness, literal plausibility, decomposability	y
Caillies (2009)	300	French	familiarity, knowledge of meaning, literality, compositionality and predictability	y
Tabossi et al. (2011)	245	Italian	knowledge, familiarity, AoA, predictability, syntactic flexibility, compositionality	y
Bonin et al. (2013)	305	French	knowledge, familiarity, subjective and objective frequency, AoA, predictability, literality, compositionality, and length	y
Citron et al. (2016)	619	German	emotional valence, arousal, familiarity, semantic transparency, figurativeness, concreteness	y
Beck and Weber (2016)	300	German	meaningfulness, familiarity, literality, decomposability;	y
Beck (2020)		English	ratings by L1 and L2 speakers	
Li et al. (2016)	350	Chinese	knowledge, familiarity, subjective frequency, AoA, predictability, literality, compositionality	y
Bulkes and Tanner (2017)	870	English	familiarity, meaningfulness, literal plausibility, decomposability, predictability	y
Nordmann and Jambazova (2017)	90	Bulgarian	familiarity, compositionality, literality, etc.	y
	100	English		
Bonin et al. (2018)	160+	French	knowledge, predictability, literality, compositionality, subjective and objective frequency, familiarity, AoA, length	y
Hubers et al. (2019)	374	Dutch	frequency of exposure, meaning familiarity, frequency of usage, transparency, imageability	y
Gavilán et al. (2021)	1252	Spanish	familiarity, knowledge, decomposability, literality, predictability, valence, arousal	y
Góngora et al. (2022)	1082	Chilean-Spanish	familiarity, ambiguity, compositionality, transparency	n
Pagliai (2023)	150	Italian	familiarity, literality, decomposability, transparency, objective knowledge, meaningfulness	y
	150	English		
Lada et al. (2024)	400	Greek	subjective frequency, ambiguity, decomposability	y
Morid and Sabourin (2024)	210	English	arousal, valence, concreteness, imageability - from English L1 and L2 speakers	y

Table 1: Published datasets from psycholinguistic research, listed chronologically. The 'Avail.' column indicates dataset availability. 'Count' indicates number of idioms.

Dataset	Authors	Language	Contents	Avail.
Phraseo-Lex	Dormeyer and Fischer (1998)	German	verbal idioms (types), with linguistic annotations	n
Berlin Idiom Project	Fellbaum and Geyken (2005)	German	500 idioms (types) with linguistic annotations	n
VNC-Tokens	Cook et al. (2008) Fazly et al. (2009)	English	53 types, almost 3000 English verb-noun combination instances annotated as to whether they are literal or idiomatic	y
OpenMWE	Hashimoto and Kawahara (2008)	Japanese	146 ambiguous idioms (types), 102846 sentences, annotated literal/idiomatic	y
IDIX	Sporleder et al. (2010)	English	78 idioms (types), 5836 instances annotated as y/n idiomatic	n
CIKB	Wang and Yu (2010)	Chinese	38K idioms (types) with linguistic annotations	n
Reddy et al. 2011	Reddy et al. (2011)	English	90 nominal compounds with compositionality ratings	y
Semeval2013 Task 5b	Korkontzelos et al. (2013)	English	85 ambiguous idioms (types), 4350 instances, each in 5-sentences context, marked y/n idiomatic	y
IDIOMENT	Williams et al. (2015)	English	580 idioms (types) with sentiment ratings; 2521 instances in sentence context, with sentiment ratings	y
Idiom Paraphrases	Pershina et al. (2015)	English	2432 idioms (types) with paraphrases, 1400 pairs of idioms annotated as y/n mutual paraphrases	y
Idiom substitution	Liu and Hwa (2016)	English	172 idioms with definitions and substitution phrases	y
IMIL	Agrawal et al. (2018)		2208 English idioms in English with their translations in seven Indian languages, 250,815 sentences with idioms	y
Idiom Translation DS	Fadaee et al. (2018)	English, German	1500 parallel sentences whose German side contains an idiom, and 1500 parallel sentences whose English side contains an idiom	y
LIdioms	Moussallem et al. (2018)	English +4	815 idioms total, in English, German, Italian, Portuguese, and Russian, with links between idioms across languages (RDF format)	y
RU idioms	Aharodnik et al. (2018)	Russian	5.4K instances of 100 idiomatic expressions (3K literal, 2.4K idiomatic), each in paragraph context.	y
CCT	Jiang et al. (2018)	Chinese	7395 Chengyu form idioms (types) and 100K context sentences	y
SLIDE	Jochim et al. (2018)	English	5000 idioms (types) with sentiment annotations	y
Senaldi 2019	Senaldi (2019)	Italian	90 verb-noun and 24 adjective-noun expressions (types)	y
Composition. of Nominal Compounds	Cordeiro et al. (2019)	English, French, Portuguese	190/180/180 nominal compounds (types), rated for compositionality	y

Table 2: Published datasets from computational linguistics research, listed chronologically. The 'Avail.' column indicates dataset availability. Table continues on the next page.

Dataset	Authors	Language	Contents	Avail.
ChID	Zheng et al. (2019)	Chinese	3848 Chengyu form idioms (types) and 518K context paragraphs with cloze blanks and multiple options	y
Swedish MWEs	Kurfalı et al. (2020)	Swedish	96 Swedish multi-word expressions (types), annotated with degree of compositionality	y
MAGPIE	Haagsma et al. (2020)	English	56622 instances of idioms in short textual context (1756 different types)	y
NCS	Garcia et al. (2021)	English, Portuguese	280 and 180 noun compounds (NCs) in English and Portuguese, with 5620/3600 sentences, marked for compositionality	y
EPIE	Saxena and Paul (2021)	English	21891 static idiom instances in sentence context (359 types) and 3135 formal idiom instances in sentence context (358 types)	y
PIE	Zhou et al. (2021)	English	823 idioms (types) with 5170 sentence-pairs containing those idioms or their literal paraphrases.	y
AStitchIn Language-Models	Tayyar Madabushi et al. (2021)	English Portuguese	223/113 nominal compounds in sentence context (4558/1872 instances), annotated as literal, idiomatic, proper noun, or 'meta usage'	y
Semeval2022 Task 2	Tayyar Madabushi et al. (2022)	English, Portuguese, Galician	5352/2555/776 instances in sentence context, extends the AStitchInLanguageModels dataset	y
PIE-English	Adewumi et al. (2022)	English	20100 samples with almost 1,200 cases of idioms (with their meanings), classified into 10 types of figurative expressions	y
FLUTE	Chakrabarty et al. (2022)	English	Part of larger dataset on figurative language. Has 1000 idiomatic sentences paired with sentences that entail or contradict the focal sentence	y
PETCI	Tang (2022)	Chinese, English	4310 Chinese idioms with 29,936 English translations.	y
ID10M	Tedeschi et al. (2022)	multiple	Training data in 10 languages was autogenerated: 10K idioms (types) 262781 sentences (32% idioms); gold test data: only for English, German, Italian Spanish, 200 sentences each	y
IDIOMEM	Haviv et al. (2023)	English	814 idiom types	y
IDEM	Prochnow et al. (2024)	English	9685 idiom-containing sentences, labelled with emotions	y
CultureLLM	Li et al. (2024a)			y
IdiomKB	Li et al. (2024b)	multiple	A merger of several datasets in English, Chinese and Japanese	y
MAPS	(Liu et al., 2024)	6 languages	proverbs and sayings, in English, German, Russian, Bengali, Chinese, Indonesian (364 to 424 per language); with entailing and non-entailing continuations	y
Multilingual Idioms and Similes in LLMs	Khoshtab et al. (2024)	Persian, English (11 total)	316 instances of idioms in LLM-generated sentence context, with entailing and non-entailing continuations. English translations provided. Also used previous figurative language datasets	y

Table 2: Continued: Published datasets from computational linguistics research.