An Idiom Benchmark for Turkish

Ebru Çavuşoğlu
Translation and Intercultural Studies
Samsun University
ebru.cavusoglu@samsun.edu.tr

Çağrı Çöltekin
Department of Linguistics
University of Tübingen
cagri.coeltekin@uni-tuebingen.de

Abstract

Despite recent significant advances, idioms, like other forms of figurative language, present a challenge to natural language processing (NLP). Benchmark corpora are essential for improving the current models on understanding idioms. However, such corpora are only available for a limited set of languages. In this paper, we introduce our ongoing work on a benchmark corpus of Turkish idioms. Our corpus is structured for testing both idiom recognition and idiom understanding. The corpus currently consists of 200 instances with sentences including idiomatic use, their literal paraphrases, similar sentences with no entailment, and non-idiomatic use of the idiomatic expressions when possible. We describe the methodology used to create the corpus, as well as initial experiments with a selection of LLMs.

1 Introduction

Idioms are multi-word expressions (MWEs) with a conventionalized interpretation. The meanings of idioms cannot be inferred from compositionality from the individual words. The correct interpretation of idioms requires familiarity with the idiom as a conventionalized unit of meaning within the particular language. Furthermore, many idiomatic expressions can also be used literally (Savary et al., 2019), leading to a possible ambiguity that has to be resolved based on the context. Similar to the other forms of figurative expression, like metaphors, proverbs, and irony, idiom understanding necessitates cultural awareness and pragmatic reasoning beyond compositional semantics because of their strong dependence on broader linguistic and non-linguistic context. As a result, idioms present challenges for non-proficient speakers, as well as the natural language processing (NLP) systems (Baldwin and Kim, 2010).

Recent developments in pretrained language models have significantly improved their performance in various tasks related to natural language generation and comprehension. However, figurative language language understanding remains to be one of the key challenges even for state-of-the-art language models (Tayyar Madabushi et al., 2021; Mi et al., 2025). Measuring and improving NLP systems beyond current state-of-the-art on figurative language processing requires high-quality and diverse benchmark datasets. However, the majority of current figurative language benchmark datasets concentrate on English or a limited number of high-resource languages. Although some multilingual idiom datasets exist (e.g., Tedeschi et al., 2022; Moussallem et al., 2018), the datasets for other languages are rather scarce.

In this paper, we present a benchmark corpus of Turkish idiomatic expressions that can be used to test idiom recognition, idiom understanding, paraphrasing, and contextual disambiguation. Each idiom in the corpus includes (1) the general form of the idiomatic expression (IE), (2) the description possibly with examples from a dictionary definition, (3) an example sentence with idiomatic use of the IE, (4) an example sentence with non-idiomatic, literal use of the IE, (5) a literal paraphrase of the idiomatic sentence (entailing (3)), and (6) a sentence with semantic/surface similarity to (3) without entailment. An example from the corpus is presented in Table 1. The fields (1) and (2) were obtained from online dictionaries, while fields (3)-(6) were created in this study. The primary objective is to provide a reliable and reusable benchmark that accurately captures linguistic variation and authentic usage of idioms in Turkish. Although multiple corpora of idiomatic expressions exist for Turkish (e.g., Berk et al., 2018; Eryiğit et al., 2023), these corpora focus on idiom detection tasks. To the best of our knowledge, a manually constructed corpus similar to our corpus does not exist for Turkish. Besides as a benchmark for

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 103–109
March 28, 2026 ©2026 Association for Computational Linguistics

assessing idiom understanding of language models, the present dataset is also a useful resource for linguistic analysis of multi-word expressions and figurative language use, and for educational applications.

In the remainder of this paper, we briefly summarize some of the earlier work in the field (Section 2), describe the methodology used during corpus creation and provide some statistics on the corpus in Section 3. In Section 4 we present results on a selection of large language models (LLMs) for idiom detection and idiom understanding tasks evaluated on the present benchmark data, before concluding in Section 5.

2 Related work

Computational study of idioms typically overlap with studies of multi-word expressions (MWEs), as well as studies that focus on figurative language. While computational models of idiom understanding have a long history, the number of studies and the number of corpora annotated for idiomatic expressions has recently grown more rapidly (see Flor et al., 2025, for a recent survey of datasets).

As in other areas of natural language processing, many influential datasets are for English (Cook et al., 2008; Liu and Hwa, 2016; Stowe et al., 2022; Chakrabarty et al., 2022; Haviv et al., 2023, e.g.,). Recently, idiom datasets for other languages, such as Korean (Wang et al., 2025) and Danish (Sørensen et al., 2025), and even for truly low-resource languages, like Nepali (Pokharel and Agrawal, 2025) and Konkani (Shaikh et al., 2024) have also been published. Another relatively recent direction is multilingual datasets like AStitchInLanguageModels (Tayyar Madabushi et al., 2021) (English and Portuguese), ID10M (Tedeschi et al., 2022) which includes 10 languages, LIdioms (Moussallem et al., 2018) which also links idiomatic expressions in the languages covered. Khoshtab et al. (2025) also unifies a number of earlier idiomatic expression datasets, as well as introducing a new one in Persian. None of these multilingual datasets include Turkish. A recent study creates a Turkish idiomatic expressions dataset (Kim et al., 2025). However, the data is not released due to copyright concerns.

There has also been a number of shared tasks with idiom-related tasks, including FigLang (Saakyan et al., 2022), and PARSEME (Ramisch et al., 2018, 2020; Savary et al., 2023) shared task. PARSEME shared task also features a Turkish MWE dataset (including idioms) which was created and improved along with the shared task (Berk et al., 2018; Ozturk et al., 2022). Besides the PARSEME data, Eryiğit et al. (2023) is another manually created idiom dataset for Turkish. Like most idiom datasets for other languages, Turkish idiom datasets so far target the idiom (span) detection task. Our work differs from these corpora as it can be used probing understanding of idiomatic expressions through entailment, paraphrasing idioms, or even for idiom generation. Furthermore, current Turkish idiomatic expression datasets typically cover a small number of potentially idiomatic expressions (with a large number of figurative/literal example sentences), while our aim is to include a large number of diverse potential idiomatic expressions.

3 Corpus Creation and Corpus statistics

We selected a large set of idioms from a number of online idiom and proverb dictionaries.[1] We removed the proverbs, based on the indication in each dictionary, and eliminated exact duplicates. This resulted in 10 970 idioms and their descriptions. Some of the descriptions also include example uses of the idiom from literature. Turkish is an agglutinative language with a wide range of inflectional and derivational morphology, as well as a flexible word order. As a result, Turkish idioms often undergo morphological changes, such as shifts in tense, person, or voice, while retaining their metaphorical meaning. For instance, the idiom *burnu sürtülmek* shows up as *burnu sürtüldü* and *burnu sürtülsün* in different examples in Table 1. The potential variation is much wider, (e.g., *sürtülmüş büyük burunları* 'their big noses are (evidentially) scraped (lit.)' can also be perfectly fine in the appropriate context).

Another variation related to the corpus creation is the potential literal use of the idiomatic expressions. Some expressions are very likely to be used in their literal meaning (e.g., *baskın yapmak* 'to raid (lit.) / to visit someone unexpectedly (fig.)'), while others are very unlikely to be used literally (e.g., *burnu havada olmak* 'to have one's nose on

[1] The dictionary of Turkish Language association (`https://sozluk.gov.tr/`, Wiktionary (`https://en.wiktionary.org/wiki/Category:Turkish_idioms`), and a Learner's dictionary of Proverbs and Idioms (`https://www.turkcedersi.net/deyimler-ve-deyimlerin-anlamlari/`).

Field	Example
Form	*burnu sürtülmek* 'to have (ones) nose scraped (lit.)'
Description	*Sıkıntı çektikten sonra daha önce beğenmediği bir durumu kabul etmek, gururundan vazgeçmek.* 'To learn a lesson, accept an (unfavorable) condition after an unpleasant experience.'
Figurative	*Sözümüzü dinlemediği için burnu sürtülsün diye bıraktık.* 'Since he/she did not listen, we left him/her there to teach him/her a lesson.'
Lit. paraphrase	*Sözümüzü dinlemediği için sıkıntı çeksin diye bıraktık.* 'Since he/she did not listen, we left him/her there for him/her to suffer (and learn).'
Similar	*Sözümü dinlemedi ve burnu büyük diye ameliyat oldu ama sonrasından sıkıntı çekti.* 'She/he did not listen to me and had a nose operation, but suffered a lot afterwards.'
Literal	*Kapıyı yüzüne birden kapatınca burnu sürtüldü.* 'When the door was shut on her/his face, his nose scraped/scratched.'

Table 1: An example from the corpus.

the air (lit.) / to be arrogant (fig.)'). All idiomatic expressions in our corpus are MWEs. Most idiomatic expressions in the corpus are verbal constructions (including nominal object/oblique modifiers) similar to ones exemplified so far (89 %). However, there are also a number of conventionalized metaphors like *boncuk gibi* 'like a bead', or other expressions like *boğazına kadar* 'up to his/her neck (lit.)' and *babasının çiftliği* 'one's fathers farm (lit.)'. Currently we do not classify the idiomatic expressions based on any of these variations.

Ideally, to have a varied benchmark, all the above-mentioned variation should be considered while selecting idioms. Unfortunately, many of these are not quantifiable. As a result, we tried to balance the frequency of the potential idiomatic expressions based on their frequency in the Leipzig web corpus (Goldhahn et al., 2012), and selecting the first 200 instances we annotate from different frequency ranges. About 30 % of the 200-idiom corpus is not observed in the corpus, while the most frequent idiomatic expression occurs 7900 times per million sentences. All 200 idiomatic forms in the current corpus occur 36 000 times per million sentences.

After selecting the 200 instances, a researcher with background in translation studies (the first author) generated sample sentences following the guidelines listed below.

- Idiomatic use of the MWE, where we aimed at natural use of the idiom in typical (informal) communication settings, where the text alone is clear enough to signal idiomatic use. We avoided the use of other idioms in the generated sentence.
- Literal paraphrase of the sentence, where the sentence with idiomatic use would entail the sentence with the literal use. We avoided paraphrasing an idiom with another idiom.
- A sentence that is similar to the sentence with the idiomatic expression, but without an entailment relation – either contradictory with the idiomatic use or irrelevant.
- Non-idiomatic use of the same MWE. Again, we avoided the use of other potentially idiomatic expressions for this sentence as well. In a few cases (3 out of 200), a non-idiomatic use did not lead to a plausible sentence (e.g., *ayağının pabucu olmak* 'to be shoe of one's feet (lit.) / to be worthless in comparison to someone (fig.)'.

The resulting corpus contains 200 idioms (797 example sentences, and dictionary descriptions). The length of the sample sentences are approximately 9 tokens on average.

4 Computational Experiments

In this section we present results of idiomaticity detection and textual entailment recognition tasks on a sample of large language models, namely Google Gemini (Gemini Team et al., 2025), OpenAI GPT 4 (OpenAI et al., 2024), and a number of smaller open models from the Llama family (Meta AI, 2024). The models are asked to perform binary classification tasks. The first task asks

Model	Detection	Entailment
Gemini 2.5-flash	0.609	0.532
GPT-4o	0.594	0.520
Llama-3 70B-Instruct	0.614	0.545
Llama-3 8B-Instruct	0.544	0.517
Llama-3 3B-Instruct	0.521	0.495
Llama-3 1B-Instruct	0.496	0.475

Table 2: Accuracy of idiom detection and entailment of a selection of LLMs on the current Turkish idiom dataset.

whether there is an idiomatic expression used figuratively in the given sentence or not. In the second task, the language model is given the idiomatic sample sentence as the premise, and either literal rephrase or semantically similar non-entailing sentence, and asked whether there is entailment or contradiction. We prompted each language model with the simple zero-shot prompts (provided in Appendix A). Prompts are given to all models in English. We experimented changing the prompting language to Turkish, and also including expressions like 'you are an expert linguist' as part of the system prompt. However, the basic prompts presented in Appendix A worked best for most cases, with some variation without clear trends. We did not experiment with few-shot or CoT prompting as our aim is not to obtain best scores, but assess the 'understanding' of the idioms by the language models without further aid, similar to what would be expected in normal language use.

Even though they were asked for a restricted set of labels, the models, especially the larger ones, occasionally offered their unsolicited reasoning. In such cases if the first or the last word still was a valid label, we used it. For a few without an identifiable label, we read the text and determined the label manually. We report accuracy as the class distribution is balanced in both tasks. Table 2 presents the accuracies of all models we experimented with in this study.

Larger models perform around 60 % of accuracy in idiomaticity detection, while smaller models perform by chance or close to chance level. There is no noticeable difference between two commercial large language models and 80B parameter Llama 3. Textual entailment scores are generally worse, again, smaller models perform at chance level. Larger models perform better than chance, but also not much better than a random baseline. Looking closely at the labels, all models seem to prefer one of the labels heavily. Larger models typically prefer the positive answer ('entailment' or 'yes' to idiomaticity), but smaller models' label preference may also vary across different runs. For the idiom instances that were not found in the Leipzig corpus, performances of large models also drop to the level of a random baseline.

5 Conclusions and Future Directions

We presented a fully manually created corpus of Turkish idioms. The corpus is built on a selection of potentially idiomatic expressions based on their frequency, and includes newly-created sample sentences including idiomatic and non-idiomatic uses of the potentially idiomatic expressions, as well as literal sentences that is in entailment or contradiction relation with the idiomatic sentence. The information in the corpus can be useful for testing idiom understanding of NLP systems through textual entailment task, paraphrasing idiomatic expressions as literal expressions, idiom generation, as well as idiom identification.

The preliminary computational experiments show that the current dataset is challenging for large language models. Even state-of-the-art commercial LLMs seem to do barely above chance level on the entailment task. We also show that the performance of the models further decreases for the low-frequency idioms. This finding is in line with earlier observations that current idiomatic expression datasets lack variety and are not challenging enough (e.g., Haagsma et al., 2019; De Luca Fornaciari et al., 2024).

The initial corpus we presented is part of an ongoing work. We plan to extend the coverage of the corpus both with respect to the number of idiomatic expressions, and with respect to sample sentences for each idiomatic expression. We also plan to classify the idiomatic expressions further, particularly the classes of expressions identified in linguistics, psychology and translation studies. These could particularly be interesting in comparing human and LM idiom usage or difficulties (e.g., attributes of idioms like 'concreteness' or 'imageability' are likely to have different difficulties for humans and LMs). The current version of the dataset is available at `https://github.com/coltekin/turkish-idioms`.

Limitations

The small size is currently the major limitation of the dataset, which also affects the reliability of the results obtained in computational experiments.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. pages 267–292.

Gözde Berk, Berna Erden, and Tunga Güngör. 2018. Turkish verbal multiword expressions corpus. In *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

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.

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)*, pages 19–22.

Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. A hard nut to crack: Idiom detection with conversational large language models. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

GülŞen Eryiğit, Ali Şentaş, and Johanna Monti. 2023. Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering*, 29(4):909–941.

Michael Flor, Xinyi Liu, and Anna Feldman. 2025. A survey of idiom datasets for psycholinguistic and computational research. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, pages 90–100, Hannover, Germany. HsH Applied Academics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 60 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Hessel Haagsma, Malvina Nissim, and Johan Bos. 2019. Casting a wide net: Robust extraction of potentially idiomatic expressions. *arXiv preprint arXiv:1911.08829*.

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.

Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2025. Comparative study of multilingual idioms and similes in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8680–8698, Abu Dhabi, UAE. Association for Computational Linguistics.

Jisu Kim, Youngwoo Shin, Uiji Hwang, Jihun Choi, Richeng Xuan, and Taeuk Kim. 2025. Memorization or reasoning? exploring the idiom understanding of LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21689–21710, Suzhou, China. Association for Computational Linguistics.

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.

Meta AI. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

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).

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 103 others. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Yagmur Ozturk, Najet Hadj Mohamed, Adam Lion-Bouton, and Agata Savary. 2022. Enhancing the PARSEME Turkish corpus of verbal multiword expressions. In *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, pages 100–104, Marseille, France. European Language Resources Association.

Rhitabrat Pokharel and Ameeta Agrawal. 2025. neDIOM: Dataset and analysis of Nepali idioms. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 160–171, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, and 6 others. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakrabarty, Debanjan Ghosh, and Smaranda Muresan. 2022. A report on the FigLang 2022 shared task on understanding figurative language. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 178–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Agata Savary, Silvio Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta, and Voula Giouli. 2019. Literal occurrences of multiword expressions: Rare birds that cause a stir. *The Prague Bulletin of Mathematical Linguistics*.

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

Nathalie Hau Sørensen, Sanni Nimb, Agnes Aggergaard Mikkelsen, and Jonas Jensen. 2025. The Danish idiom dataset: A collection of 1000 Danish idioms and fixed expressions. In *Proceedings of the 1st Workshop on Nordic-Baltic Responsible Evaluation and Alignment of Language Models (NB-REAL 2025)*, pages 55–63, Tallinn, Estonia. The University of Tartu Library.

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

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.

Xiaonan Wang, Seoyoon Park, and Hansaem Kim. 2025. Benchmarking Korean idiom understanding: A comparative analysis of local and global models. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 1341–1351, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

A Prompts

The following are the prompts used for the experiments reported in the paper.

```
I will provide you with a
pair of sentences in Turkish
consisting of a premise and
a hypothesis. Is there a
contradiction or entailment
between the premise and
hypothesis? Answer only with
"contradiction" or "entailment".
Premise: [P]
Hypothesis: [H]
Label:
```

```
Does the following Turkish
sentence contain an idiom which
is used figuratively? Answer
only with "yes" or "no".
Sentence: [S]
Answer:
```