METAPHORSHARE: A Dynamic Collaborative Repository of Open Metaphor Datasets

Joanne Boisson, Arif Mehmood and Jose Camacho-Collados
Cardiff NLP, School of Computer Science and Informatics
Cardiff University, United Kingdom
{boissonjc,mehmooda3,camachocolladosj}@cardiff.ac.uk

Abstract

The metaphor studies community has developed numerous valuable labelled corpora in various languages over the years. Many of these resources are not only unknown to the NLP community, but are also often not easily shared among the researchers. Both in human sciences and in NLP, researchers could benefit from a centralised database of labelled resources, easily accessible and unified under an identical format. To facilitate this, we present METAPHORSHARE, a website to integrate metaphor datasets making them open and accessible. With this effort, our aim is to encourage researchers to share and upload more datasets in any language in order to facilitate metaphor studies and the development of future metaphor processing NLP systems. The website has four main functionalities: upload, download, search and label metaphor datasets. It is accessible at www.metaphorshare.com.

1 Introduction

The topic of figurative language processing has been addressed since early years of artificial intelligence research (Martin, 1990; Fass, 1997; Kintsch, 2000), inspired by philosophical (Richards, 1936), linguistic (Wilks, 1973) and then cognitive science theories (Lakoff and Johnson, 1980) and experimental studies (Katz et al., 1988). In spite of the influence of the different disciplines on one another, the resources created to train and evaluate NLP models have often been distinct from the resources created for metaphor studies in other fields, with the notable exception of Master Metaphor List (Lakoff, 1994) and the VU Amsterdam Metaphor Corpus (Steen, 2010).

In recent years, metaphor processing has attracted more and more attention with the progress made possible by transformer-based language models, for example Mao et al. (2019) or Zeng and Bhat (2021), and in particular the large LMs such as GPT-3 (Brown et al., 2020) used by Wachowiak and Gromann (2023) for metaphor identification. Several dedicated workshops with shared tasks (Ghosh et al., 2022, 2024; Sharma et al., 2020; Tayyar Madabushi et al., 2022a) have been organized, leading to the creation of more resources. The proper handling of figurative language by models is of crucial importance for improving performance in downstream tasks (Han et al., 2022; Li et al., 2024).

From the perspective of Humanities, tools for searching corpora such as Sketch Engine (Kilgarriff et al., 2014), annotation tools (Koller et al., 2008) and data analysis softwares have become part of the standard corpus-based research methodology. Effective adaptive support for metaphor identification and interpretation is the next step of this on-going inter-disciplinary collaboration. There were two limitations until recently: metaphor identification systems were not accurate enough to be used on free text, and different researchers are looking for different metaphors.

On one hand, numerous datasets labelled with metaphoric usage of words are created by researchers in many languages every year, and remain out of the scope of NLP research, tool development, and evaluations of the systems. This happens because most datasets are published in non NLP venues, they might not be publicly available, or might be tagged in an under-specified format that does not allow direct automatic processing (e.g. tagging precisely a metaphoric expression within a text). On the other hand, metaphor studies researchers often work on domain specific corpora, languages other than English - for which NLP has developed the most resources - and use different definitions or retain different types of metaphors for different projects, with no convenient way to systematically compare their labels with other researchers, or to rely on an existing metaphor automatic labelling tool that answers their project specific need.

We propose to help speeding up collabora-

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations), pages 509–521

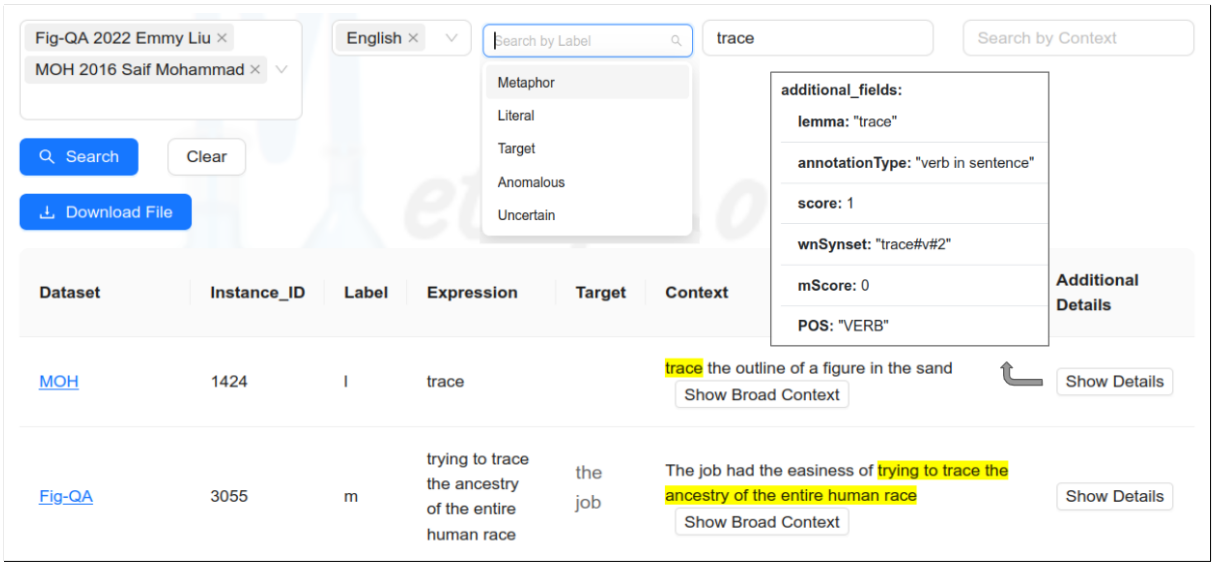

Figure 1: METAPHORSHARE search page. Specific datasets, languages, and tag types can be selected, and a text-based search within tagged expressions or into the entire text is implemented. Additional features provided with the record appear when clicking the *Show Details* button.

tions between AI/NLP communities and linguistics/metaphor studies by facilitating the unification of dataset formats and the access to existing resources by everyone, while preserving the information encoded in the original datasets. MetaphorShare is a website designed at sharing new and publicly available existing resources (see Figure 1). It is an online dynamic repository where anyone can upload or download open metaphor datasets. Our online labeling tool also offers the possibility to directly create a dataset, and optionally share it. Our hope is that the website will create a new synergy between different components of metaphor studies, limit the loss of datasets, catalog them, and facilitate their automatic processing through a unified format. Our target audience comprises anyone conducting corpus-based research on metaphors, including NLP researchers for the development of metaphor processing and identification systems[1].

Recent advances in metaphor processing are presented further in Section 2. The diversity of existing resources and the unified dataset input format is presented Section 3. The website is structured around four main functionalities that are described in Section 4: upload, download, search and label metaphor datasets. Section 5 shows the potential of our framework for facilitating NLP experiments, by evaluating a RoBERTa (Liu et al., 2019) model on a cross-dataset classification task.

[1]A short introduction to METAPHORSHARE is available at https://youtu.be/Fi48SOjueEE

2 Related work

Who studies and labels metaphors? Cameron and Maslen (2010) summarizes the landscape of metaphor studies. The core role played by Lakoff and Johnson (1980)'s Conceptual Metaphor Theory (CMT) in Cognitive Science had an impact on the methodology developed for the analysis of metaphors in Linguistics and Literature. Metaphor analysis also became a standard approach in Anthropology, Educational Research, Political Science, or Management Research. The purpose of metaphor analysis in these disciplines is to uncover latent meaning present in discourse on a studied topic, that is sometimes conveyed elusively in corpora. Recently, Baleato Rodríguez et al. (2023) introduced propaganda modeling, integrating metaphor identification to the automatic detection of persuasive intentions.

Metaphor related websites. Veale and Li (2012) develop a system of metaphor interpretation and generation, METAPHOR MAGNET, that relies on the harvest of stereotypes from Google n-grams. The stereotypical attributes associated with concepts are leveraged for the suggestion of relevant metaphors and their interpretation (e.g. given *Google is –Microsoft*, the system outputs *giant* with properties like *lumbering* and *sprawling*). More recently, Mao et al. (2023) release an end-to-end domain independent metaphor identification and interpretation website for free text: METAPRO tags

metaphoric tokens or multi-word expressions in sentences, generates paraphrases, and outputs a more abstract concept mapping derived from the analysis of the sentence. The authors list limitations due to the lack of training data for inference and extended metaphors. The main difference with these two initiatives and ours is that our aim is to integrate already labeled datasets into a unified repository, encouraging the structuring of resources and expression of needs from the metaphor study community, before building automatic annotation tools. Similarly to ours, MetaPro also uses several metaphor datasets for training the models (Mao et al., 2019), but in this case they do not mention more than six datasets, all of them often used by the NLP community.

Annotation tools. Many open source (e.g. Label Studio, Dataturks, Doccano (Nakayama et al., 2018) and Potato (Pei et al., 2022)) and proprietary (e.g. LabelBox, Prodigy or Amazon Mechanical Turk) annotation tools exist. Our purpose is not to compete with these services but rather to offer a simple targeted experience for metaphor annotation, in which the output is directly mapped to the unified repository format. This way, non-expert users can avoid the need for learning a general tool, as well as having to learn how to modify given labeled outputs, which would not be trivial for most.

Metaphor processing in NLP. One important usage of our website is to facilitate the creation of personalised metaphor processing models. Metaphor processing in NLP comprises many methods developed for metaphor identification (Turney et al., 2011; Tsvetkov et al., 2014; Mao et al., 2019; Wachowiak and Gromann, 2023), but also generation (Veale, 2016; Stowe et al., 2021; Chakrabarty et al., 2021b), textual (Mao et al., 2018) and multimodal (Kulkarni et al., 2024) interpretation, metaphor understanding through entailment (Agerri et al., 2008; Chakrabarty et al., 2021a; Stowe et al., 2022), among other tasks. Ge et al. (2023) provide a comprehensive recent survey on the topic.

3 Data Sharing

METAPHORSHARE aims at facilitating computational research on multiple datasets. A minimally constrained format is required to store the datasets into a database, conveniently compare the resources, search efficiently into the different fields across datasets, and run experiments on several sets at once. Metaphor datasets can be added by their authors or by anyone else if their license allows it. In this section, we review the existing dataset formats and our unification choices. The website database currently contains 25 datasets. We select twelve datasets with open licenses and diverse encoded information to illustrate our unification process and as representative examples for our repository. Their characteristics are summarized in Table 1 and in more details in the appendix (Table 4).

Name	Reference	N	%M
Words in syntactically constrained sentences (Psycholinguistics)			
JANK	Jankowiak (2020)	240	50
CARD_V	Cardillo et al. (2010)	280	50
CARD_N	Cardillo et al. (2010, 2017)	512	50
Words in natural short contexts (NLP)			
MOH	Mohammad et al. (2016)	1632	25
NewsMet	Joseph et al. (2023)	1205	49
TSV_A	Tsvetkov et al. (2014)	1945	50
GUT	Gutiérrez et al. (2016)	8591	54
MWE in natural long context (NLP)			
PVC	Tu and Roth (2012)	1348	65
MAD	Tayyar Madabushi et al. (2022b)	4558	48
MAGPIE	Haagsma et al. (2020)	48395	75
Words sampled from VUAC (MIPVU)			
VUAC_BO	Boisson et al. (2023)	39223	52
TONG	Tong et al. (2024)	1428	46

Table 1: Twelve metaphor datasets included in the METAPHORSHARE repository, with a single metaphor/literal tag per entry. Short contexts consist of one word or a short sentence. Long contexts correspond to multiple sentences. Datasets derived from VUAC have a document level context. The N column shows the number of instances in the datasets. %M indicates the percentage of expressions labelled as metaphorical.

One tagged expression per example. Many datasets in NLP contain labels for binary classification of one delimited expression within a given context. The expressions considered might be single tokens, multi-word expressions (MWE), phrases or compounds. The task of the original data release paper may be to decide weather the marked expression is used metaphorically or literally, or to find a correct literal paraphrase of the metaphoric expression, among other possible analysis and tasks. We define a pair of constrained tags <*m*> and <*l*> to mark metaphoric of literal annotated expressions in text. Some datasets, e.g. Jankowiak (2020), also contain anomalous sentences, that are handled with an additional <*a*> tag.

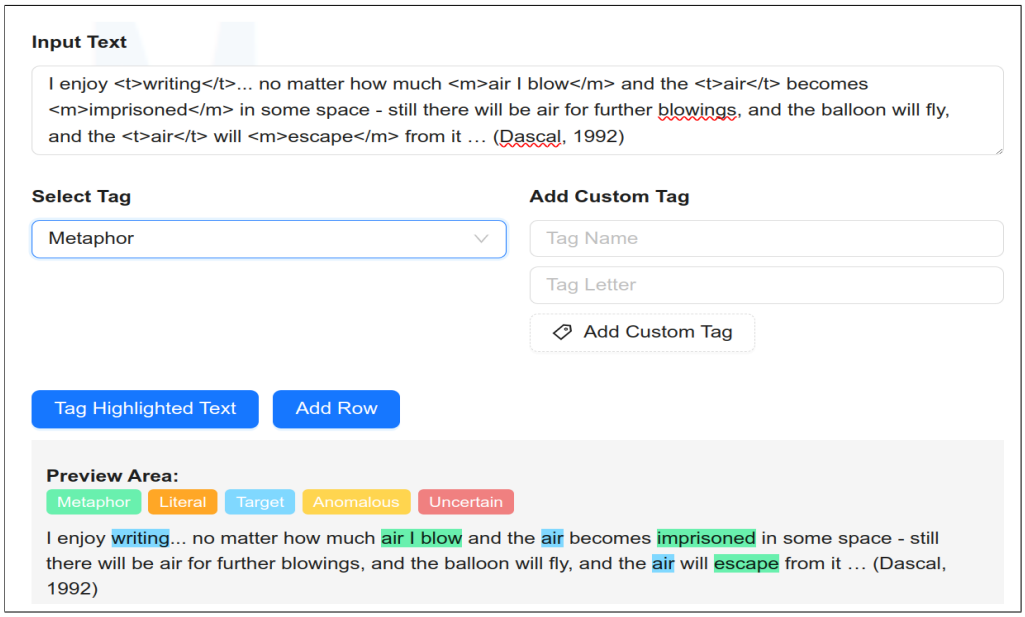

Figure 2: Screenshot of the online annotation tool showing the text input area, tag selection and creation, and resulting tagged text highlighted in different colours.

Multiple tags per example. Steen (2010) defined an annotation guideline that is widely used in metaphor studies, the MIPVU procedure, together with the release of the VU Amsterdam Corpus (VUAC), the largest existing metaphor corpus. Other examples of English open datasets created following MIPVU, that are in METAPHORSHARE, are Nacey et al. (2019b) and Julich-Warpakowski (2022) on the Music Criticism domain. Many other MIPVU datasets exist in various languages with adaptations of the guideline (Nacey et al., 2019a), for example (Nacey, 2022) for Norwegian. In this framework, every token in a document is labelled as metaphoric or literal.

Different versions of the VUAC suiting NLP tasks, adapted for binary classification tasks (Leong et al., 2018), or including additional ratings of novelty (Parde and Nielsen, 2018) have been published over the years. In the evaluation section, Section 5, we use a modified version of the VUAC corpus, adapted for binary classification (VUAC_BO) as described in Boisson et al. (2023). METAPHORSHARE accepts entries with multiple tags per text span, as exemplified in Figure 2. MIPVU datasets are integrated in our repository with two versions fitting different possible NLP tasks: multiple tags per sentence, and one tag per sentence, with duplicated sentences.

Context. The context of a tagged expression may consist of one single word, in the case of adjective-noun pairs datasets (Gutiérrez et al., 2016; Tsvetkov et al., 2014). It often consist of one sentence such as Mohammad et al. (2016) or Cardillo et al. (2010, 2017) or Jankowiak (2020). Longer contexts are sometimes directly provided, such as in Tayyar Madabushi et al. (2022b) or Haagsma et al. (2020) who include two sentences before and after the sentence containing the marked Potential Idiomatic Expression (PIE), or indirectly provided with a reference to the original document, such as for example in Joseph et al. (2023) who share references to newspaper articles or Tu and Roth (2012) who link the annotated examples to sentence pointers in the British National Corpus. We design options to distinguish and display narrow and broad contexts surrounding a tagged expression.

Additional Information. Beyond metaphorical/literal labels of expressions in context, metaphor datasets may contain additional information. Some datasets, for example the LCC (Mohler et al., 2016) and Gordon et al. (2015), tag the fragment of the context that carries the lexical information about the target of a metaphor (e.g. in the phrase *ocean of happiness*, the metaphoric expression is *ocean* and the target lexical cue of the metaphor is *happiness*). We define a specific tag <*t*> for such lexical cues and a free tag <*u*> for less frequently used in-text annotations.

Additional information might also be contin-

uous or integer variables such as metaphoricity/figurativeness ratings averages (Cardillo et al., 2010; Dunn, 2014; Katz et al., 1988), confidence scores, concreteness scores, novelty scores (Parde and Nielsen, 2018), level of emotion scores (Mohammad et al., 2016), frequency in corpora... They may also contain categorical variables such as PoS, textual source and genre, metaphor type, source and target concept/domains information (Gordon et al., 2015). Figure 1 shows an example of the additional information encoded in the Mohammad et al. (2016) dataset for one instance, as diplayed on METAPHORSHARE. Such features are unified through name recommendations in the guideline.

Unified input format. METAPHORSHARE currently accepts CSV files having minimally one column named *tagged_text*. It must contain text with at least one tagged expression per line. The accepted tag set is: *< m: metaphoric, l: literal, t: target, a:semantically anomalous, u: free tag>*. Multiple tags per entry are allowed , e.g. *I <m>swim</m> today in an <m>ocean</m> of <t>happiness</t>*. Additional information may be added in freely named columns. Users are encouraged to follow naming recommendations for common features. Sentence indices can be added to preserve the sentence order in a document and handle datasets labeled at the discourse level. All the datasets containing examples in languages other than English could in theory be uploaded in the flexible format defined above, e.g. Polish sentences in Jankowiak (2020), Mandarin Chinese sentences in Wang (2022), Farsi sentences in Levin et al. (2014) and Norwegian examples in Nacey (2022) can be converted smoothly. The free tag *<u>*, open features defined in the data file, and built-in multilingual text search functionalities of the METAPHORSHARE Elasticsearch database make our framework adaptable to the specificities of a language. Moreover, we will provide individual support. If any author of a dataset encounters issues reformatting a dataset into our expected format, we invite them to contact us by email for support.

4 MetaphorShare: The Website

MetaphorShare general architecture is described in Section 4.1. The website is organized around four main pages: an uploading page (Section 4.2), a dataset catalog (Section 4.3), a page for searching into the datasets records (Section 4.4), and an online annotation tool (Section 4.5).

Field	Description
Required	
Name	Name of the person uploading data
Email	Address of the person uploading data
Dataset	Name of the dataset
Author	Main author of the dataset
CSV file	Formatted dataset
License	New or existing dataset license
Optional	
Paper title	Publication presenting the dataset
Author(s)	Authors of the paper
Year	Paper's publication year
Reference	Bibtex
Languages	Languages of the labelled examples
Source	Source corpora of the labelled instances
Genre	Novel, poetry, news, spoken language...
Target POS	PoS of the target if any
Source POS	PoS of the source expression if any
Annot. num.	Number of annotators
Annot. profile	Linguists, authors, crowdsourcing...
IAA	Inter-annot. agreement metric & score
Comments	Additional description and comments

Table 2: Required and optional dataset information to fill when uploading a new dataset

4.1 Website architecture

MetaphorShare is hosted by Cardiff University School of Computer Science and Informatics. It employs SSL protocols and HTTPS to safeguard data transmission. Direct access to the database is restricted, with backend schedulers managing all search and upload operations, thus fortifying data protection.

Back & front end. The backend is developed with Python FastAPI, chosen for its speed and ease of use. It features schedulers for routine tasks such as data ingestion and file cleanup. Integration with Elasticsearch allows for efficient indexing and retrieval of dataset entries. Additionally, dataset metadata is securely stored in a PostgreSQL database, ensuring data integrity and structured storage. The frontend is developed using ReactJS (version 18), leveraging libraries such as Bootstrap and Ant Design. State management is handled by Redux, and data visualization is powered by Chart.js.

Database. Our system employs a dual-database approach to efficiently manage and query data. A PostgreSQL database serves as the primary storage for user-submitted data during the dataset upload process, and stores them ensuring data integrity and facilitating efficient data management. An Elastic-

Field	Description
Required	
id	Database record ID
dataset_id	Link to the Dataset index
expression	Token or expression being labelled
label	Metaphoric, literal, anomalous, other
position	Expression offset in the given context
Optional	(often present)
sent_index	Index preserving the sentences order
reference	Reference to a source document
pos	Part of Speech
long_context	Context surrounding a labelled sentence
target	Lexical cue of the target concept
target_position	target offset in the given context
source	Metaphor source concept/domain
target	Metaphor target concept/domain
mscore	Metaphoricity/figurativeness score
Free	Any field, for example *emotion ratings*

Table 3: Fields to describe an instance of a dataset in the ElasticSearch database

search engine enhances the search capabilities of METAPHORSHARE with exact and fuzzy match for text field supported in multiple languages.

4.2 Uploading a new dataset

Once a dataset is formatted into a CSV file, as described in Section 3, and once its license is specified by the authors, the dataset can be uploaded by any user.

Dataset metadata. The required and optional fields to be filled in at uploading time are presented in Table 2. For example, if a dataset is associated with one or several publications, the references should be added in the appropriate fields, preferably in BIBTEX format. The information provided will be displayed in the dataset catalog (c.f. Figure 4 in the appendix), and will help to further categorize the datasets in the future versions of the website.

Automatic and manual validations. After validating the form, an immediate automatic file format check on the data is done and feedback is returned to the user for acceptance or rejection of the file. Outputs of this step can be seen in Figure 5 & 6 in the appendix. In a second manual validation step, the website administrators check the license information, dataset reference, and eventually suggest modifications of CSV free columns names, that contain the dataset specific additional information, in an email interaction with the user, for database field unification purposes.

Storage in the database. Once a dataset passes the two validation steps, it is stored in the main repository database. This ElasticSearch database has two indices: the *Datasets* index similar to Table 2 fields stores all the user provided metadata, and information extracted during the processing of the data file such as the dataset size and the label distribution. The second one, the Potentially Metaphoric Expressions, *PME* index, contains the output of the data file parse, as shown Table 3. Each textual labelled example is linked to the *Dataset* index.

4.3 Dataset Catalog

The purpose of this page is to display the main features of each dataset included in the repository. The fields filled at uploading time, including the license chosen, are displayed in the dataset catalog page. Statistics computed from the input file are also shown and plotted, in particular the number of instances, label set and distribution. The list of additional fields is also extracted from the input file. A partial view of a dataset summary in the catalog is shown Figure 4 in the appendix.

4.4 Search Page

The records can be accessed and downloaded through a search page, as shown in Figure 1. The search fields include optional selections of datasets and languages from drop-down lists, restricted search for a label (e.g. *metaphorical* or *literal*), and keyword search into the PME text (with exact and fuzzy match), or into the entire text. One phrase or sentence is shown by default on the result page, but longer context can be accessed in one click, as well as all the additional information attached to a record.[2] The result of any search can be downloaded as a CSV file.

4.5 Online data annotation

On the Edit/Label Dataset page, one can choose to directly input some text into the interface of the annotation tool, or alternatively to upload a table (in a CSV or Excel format) containing unannotated or partially annotated text. An input can then be labelled from scratch or edited by pairing highlighted expressions with tags (c.f. Figure 2). While the METAPHORSHARE tagset is predefined, custom tags can also be created. The resulting labelled data (c.f. Figure 7 in the appendix) can

[2]The organization of the catalog and the fine-grained search options are likely to evolve in future versions of METAPHORSHARE.

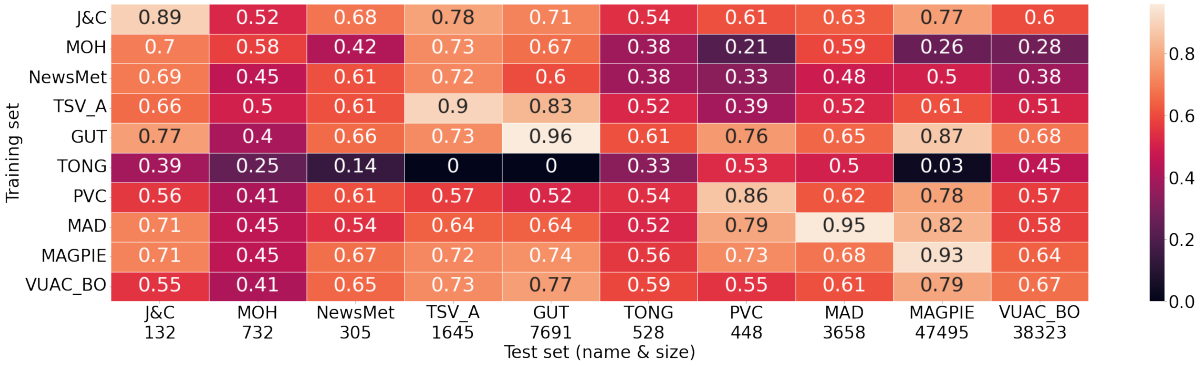

Figure 3: Results of the cross dataset evaluation. F1-score of the *metaphor* class. Each training set contains 800 examples and the test sets sizes are shown on the x axis.

then be downloaded as a CSV, or directly uploaded in METAPHORSHARE following the procedure described in Section 4.2. This online annotation functionality facilitates the creation of resources immediately compatible with our accepted formats, and speeds up the correction of CSV files that do not pass the automatic validation step of the dataset upload procedure. It is also an accessible tool for annotators without any computer science background.

5 Evaluation

As a sanity check, we first verified that METAPHORSHARE uploading, database insertion and search functionalities are working properly with the datasets presented in Table 1.

Then, as a case study we perform a cross-dataset metaphor identification analysis to illustrate a possible usage of METAPHORSHARE for NLP research. Given an input expression, the task consists of deciding whether the expression is a metaphor or not. Different projects label metaphors for different purposes. METAPHORSHARE makes it easy to fine-tune models on specific datasets to further support project-specific automatic or semi-automatic labeling.

Experimental setting. In order to maintain a training set of the same size for all datasets, we randomly sampled 800 examples from each set to create the training data. Because our datasets from psycholinguisics are small and similar in their creation method, we grouped them into one single set called J&C in this experiment. RoBERTa base models (Liu et al., 2019) are then finetuned independently on the 10 datasets and tested on each of them. Similarly to the experiments in Boisson et al. (2023), for hyperparameter optimisation, we rely on the Bayesian Optimization with the Hyperband (BOHB) algorithm (Falkner et al., 2018) available in RayTune (Liaw et al., 2018), with an identical hyperparameter search space and 25 trials.

Results. Results are shown in Figure 3. As expected, models trained and evaluated on the same dataset often achieved the best results. A few datasets generalise better than others. It is surprisingly the case for J&C and GUT, both containing only short example with constrained syntactic structures. On the other hand, datasets such as TONG and NEWS do not generalize as well, probably due to different labelled PoS, to the token or MWE span of the labels, or due to redundant contexts corresponding to datasets created for other tasks such as paraphrasing in the case of TONG.

6 Conclusion and Future Work

In this paper, we have presented a website to label, unify and share metaphor datasets. It enables an easy integration of new resources of different original formats by the community. There are currently 25 English datasets integrated in the platform. In the future, we are planning to integrate and reach communities working on other languages as one of the main aims.

As far as the platform is concerned, we will focus on integrating a functionality for automatically labelling raw text. The progress recently made by transformer-based language models for metaphor identification opens the path to the creation of helpful personalized tools for automatic or semi-automatic labelling of metaphors based on an initial sample of annotations.

Acknowledgments

This research was supported by the Association for Researching and Applying Metaphors through the *Building Bridges* fund. We thank the British National Corpus for giving us the permission to share fragments of texts of the BNC directly on MetaphorShare. We also thank Eileen Cardillo, all the authors of Gutiérrez et al. (2016) and of Birke and Sarkar (2006) for allowing us to redistribute their datasets online for research purposes, and to the rest of the authors of the datasets included in the repository for the creation of metaphor datasets with a flexible open license. Without the effort of the metaphor community, this work would not have been possible. Jose Camacho-Collados was supported by a UKRI Future Leaders Fellowship.

References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Rodrigo Agerri, John Barnden, Mark Lee, and Alan Wallington. 2008. Textual entailment as an evaluation framework for metaphor resolution: A proposal. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, pages 357–363. College Publications.

Daniel Baleato Rodríguez, Verna Dankers, Preslav Nakov, and Ekaterina Shutova. 2023. Paper bullets: Modeling propaganda with the help of metaphor. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 472–489, Dubrovnik, Croatia. Association for Computational Linguistics.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the 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*, pages 6581–6590, Singapore. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the Annual Conference on Neural Information Processing Systems*.

L. Cameron and R. Maslen. 2010. *Metaphor Analysis: Research Practice in Applied Linguistics, Social Sciences and the Humanities*. Studies in applied linguistics. Equinox.

Eileen R. Cardillo, Christine Watson, and Anjan Chatterjee. 2017. Stimulus needs are a moving target: 240 additional matched literal and metaphorical sentences for testing neural hypotheses about metaphor. *Behavior Research Methods*, 49(2):471–483.

E.R. Cardillo, G.L. Schmidt, A. Kranjec, and A. Chatterjee. 2010. Stimulus design is an obstacle course: 560 matched literal and metaphorical sentences for testing neural hypotheses about metaphor.

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

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021b. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261, Online. Association for Computational Linguistics.

Jonathan Dunn. 2014. Measuring metaphoricity. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 745–751, Baltimore, Maryland. Association for Computational Linguistics.

Stefan Falkner, Aaron Klein, and Frank Hutter. 2018. Bohb: Robust and efficient hyperparameter optimization at scale. *Preprint*, arXiv:1807.01774.

Dan Fass. 1997. Processing metaphor and metonymy.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukwac , a very large web-derived corpus of english.

Mengshi Ge, Rui Mao, and Erik Cambria. 2023. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*, 56(02):1829–1895.

Debanjan Ghosh, Beata Beigman Klebanov, Smaranda Muresan, Anna Feldman, Soujanya Poria, and Tuhin Chakrabarty, editors. 2022. *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid).

Debanjan Ghosh, Smaranda Muresan, Anna Feldman, Tuhin Chakrabarty, and Emmy Liu, editors. 2024. *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*. Association for Computational Linguistics, Mexico City, Mexico (Hybrid).

Jonathan Gordon, Jerry Hobbs, Jonathan May, Michael Mohler, Fabrizio Morbini, Bryan Rink, Marc Tomlinson, and Suzanne Wertheim. 2015. A corpus of rich metaphor annotation. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 56–66, Denver, Colorado. Association for Computational Linguistics.

E. Dario Gutiérrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin Bergen. 2016. Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 183–193, Berlin, Germany. Association for Computational Linguistics.

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

Sooji Han, Rui Mao, and Erik Cambria. 2022. Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 94–104, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Katarzyna Jankowiak. 2020. Normative data for novel nominal metaphors, novel similes, literal, and anomalous utterances in polish and english. *Journal of Psycholinguistic Research*, 49(4):541–569.

Rohan Joseph, Timothy Liu, Aik Beng Ng, Simon See, and Sunny Rai. 2023. NewsMet : A 'do it all' dataset of contemporary metaphors in news headlines. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10090–10104, Toronto, Canada. Association for Computational Linguistics.

Nina Julich-Warpakowski. 2022. *Motion Metaphors in Music Criticism: An empirical investigation of their conceptual motivation and their metaphoricity*. Metaphor in Language, Cognition, and Communication. John Benjamins Publishing Company.

Albert Katz, Allan Paivio, Marc Marschark, and Jim Clark. 1988. Norms for 204 literary and 260 nonliterary metaphors on 10 psychological dimensions. *Metaphor and Symbol - METAPHOR SYMB*, 3:191–214.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1:7–36.

Walter Kintsch. 2000. Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, 7(2):257–266.

Veronika Koller, Andrew Hardie, Paul Rayson, Elena Semino, and Lancaster. 2008. Using a semantic annotation tool for the analysis of metaphor in discourse. *Metaphorik.De*, 15.

Shreyas Kulkarni, Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2024. A report on the FigLang 2024 shared task on multimodal figurative language. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 115–119, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

G. Lakoff. 1994. *Master Metaphor List*. University of California.

George Lakoff and Mark Johnson. 1980. *Metaphors we Live by*. University of Chicago Press, Chicago.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans, Louisiana. Association for Computational Linguistics.

Lori Levin, Teruko Mitamura, Brian MacWhinney, Davida Fromm, Jaime Carbonell, Weston Feely, Robert Frederking, Anatole Gershman, and Carlos Ramirez. 2014. Resources for the detection of conventionalized metaphors in four languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 498–501, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024. Finding challenging metaphors that confuse pretrained language models. *Preprint*, arXiv:2401.16012.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E. Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training. *Preprint*, arXiv:1807.05118.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rui Mao, Xiao Li, Kai He, Mengshi Ge, and Erik Cambria. 2023. MetaPro online: A computational metaphor processing online system. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 127–135, Toronto, Canada. Association for Computational Linguistics.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and WordNet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1222–1231, Melbourne, Australia. Association for Computational Linguistics.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898, Florence, Italy. Association for Computational Linguistics.

James H. Martin. 1990. *A Computational Model of Metaphor Interpretation*. Academic Press Professional, Inc., San Diego, CA, USA.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

Susan Nacey. 2022. Replication data for: Systematic metaphors in Norwegian doctoral dissertation acknowledgements.

Susan Nacey, Aletta G. Dorst, Tina Krennmayr, and W. Gudrun Reijnierse, editors. 2019a. *Metaphor Identification in Multiple Languages: MIPVU around the world*. John Benjamins.

Susan Nacey, Tina Krennmayr, Aletta G. Dorst, and W. Gudrun Reijnierse. 2019b. Replication Data for: What the MIPVU protocol doesn't tell you (even though it really does).

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Natalie Parde and Rodney Nielsen. 2018. Exploring the terrain of metaphor novelty: A regression-based approach for automatically scoring metaphors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

I.A. Richards. 1936. *The Philosophy of Rhetoric*. Bryn Mawr College. Mary Flexner lectures. Oxford University Press.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Gerard Steen. 2010. *A method for linguistic metaphor identification: from MIP to MIPVU*, volume v. 14 of *Converging evidence in language and communication research*. John Benjamins Pub. Co., Amsterdam.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6724–6736, Online. Association for Computational Linguistics.

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, 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)*. Association for Computational Linguistics.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for llms. *Preprint*, arXiv:2403.11810.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Yuancheng Tu and Dan Roth. 2012. Sorting out the most confusing english phrasal verbs. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Tony Veale. 2016. Round up the usual suspects: Knowledge-based metaphor generation. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 34–41, San Diego, California. Association for Computational Linguistics.

Tony Veale and Guofu Li. 2012. Specifying viewpoint and information need with affective metaphors: A system demonstration of the metaphor-magnet web app/service. In *Proceedings of the ACL 2012 System Demonstrations*, pages 7–12, Jeju Island, Korea. Association for Computational Linguistics.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032, Toronto, Canada. Association for Computational Linguistics.

Xin Wang. 2022. Normed dataset for novel metaphors, novel similes, literal and anomalous sentences in chinese. *Frontiers in Psychology*, 13.

Yorick Wilks. 1973. Preference semantics.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

A Detailed information of datasets used in our experiments

Table 4 provides detailed information about the datasets, including license, source corpora, label and sentence distribution. Additional links and preprocessing details are listed below :

JANK: Anomalous sentences and simile are not shown in the table. because because they are not used in our binary classification evaluation task .

MOH: The license is available at `https://saifmohammad.com/WebPages/metaphor.html`. The original dataset contains 1639 instances. A few duplicated example sentences caused by orthographic variants of the target word, such as distil/distill have been removed.

NewsMet: We show the manually labelled sentences (named *gold* by the authors). The corpus from which the sentences are extracted can be found at `https://github.com/several27/FakeNewsCorpus`.

TSVET_A: The license can be found at `https://github.com/ytsvetko/metaphor/blob/master/LICENSE.md`

GUT: The UKaC is presented in Ferraresi et al. (2008).

PVC: The original dataset is accessible at `https://cogcomp.seas.upenn.edu/page/resource_view/26`

MAGPIE: The PMB corpus from which some sentences are sourced is presented in Abzianidze et al. (2017).

TONG: Original VUAC sentences and *apt* (dataset new label) paraphrases are counted in the table.

B Dataset upload

Automatic file format check. Figure 5 & 6 show the row by row feedback provided after the automatic validation step when the file is rejected, and the pop up message indicating that a file passed the automatic validation step.

Dataset information after validation. Figure 4 shows a part of a dataset presentation page once it is integrated into the catalog and database.

C Dataset catalog

Figure 4 is a screenshot of the top of the page presenting a dataset in the Catalog section of METAPHORSHARE. The dataset name, main author and research paper associated to the data release are displayed.

D Online annotation tool

Figure 7 shows how the created records of a dataset are displayed, allowing an easy step by step annotation.

Dataset Name	Reference	License	N	N dist. ctxt.	N dist. expr.	% met.	Expr. PoS	Domain/Source
Words in syntactically constrained sentences (Psycholinguistics)								
JANK	Jankowiak (2020)	CC BY 4.0	240	240	120	50	N	constructed examples
CARD_V	Cardillo et al. (2010)	CC BY-NC	280	280	140	50	V	constructed examples
CARD_N	Cardillo et al. (2010, 2017)	CC BY-NC	512	512	256	50	N	constructed examples
Words in natural short contexts (NLP)								
MOH	Mohammad et al. (2016)	see data page	1632	1632	439	25	V	WordNet examples
NewsMet	Joseph et al. (2023)	Apache-2.0	1205	1205	477	49	V	Fake News Corpus
TSV_A	Tsvetkov et al. (2014)	see data page	1945	1072	687	50	A	various websites
GUT	Gutiérrez et al. (2016)	AFL-3.0	8591	3479	23	54	A	Wikipedia, UKWaC...
Multi-Word Expressions in natural long contexts (NLP)								
PVC	Tu and Roth (2012)	no license	1348	1348	23	65	V-Prep	BNC
MAD	Tayyar Madabushi et al. (2022b)	GPL-3.0	4558	4554	251	48	NC	Common Crawl
MAGPIE	Haagsma et al. (2020)	CC-BY-4.0	48395	47283	9307	75	Various	BNC, PMB.
Words sampled from VUAC (MIPVU)								
VUAC_BO	Boisson et al. (2023)	CC BY-SA 3.0	39223	11476	8674	52	Various	VUAC
TONG	Tong et al. (2024)	CC-BY-4.0	1428	739	861	46	Various	VUAC & paraphrases

Table 4: Twelve example metaphor datasets included in the METAPHORSHARE repository. The N column shows the number of instances in the datasets, followed by the number of distinct provided context surrounding the potential metaphoric expression (N dist ctxt), and the number of distinct lemmas labelled metaphoric or literal (N dist. expr.). %met. indicates the percentage of expressions labelled as metaphorical, and Expr. PoS their part of speech.

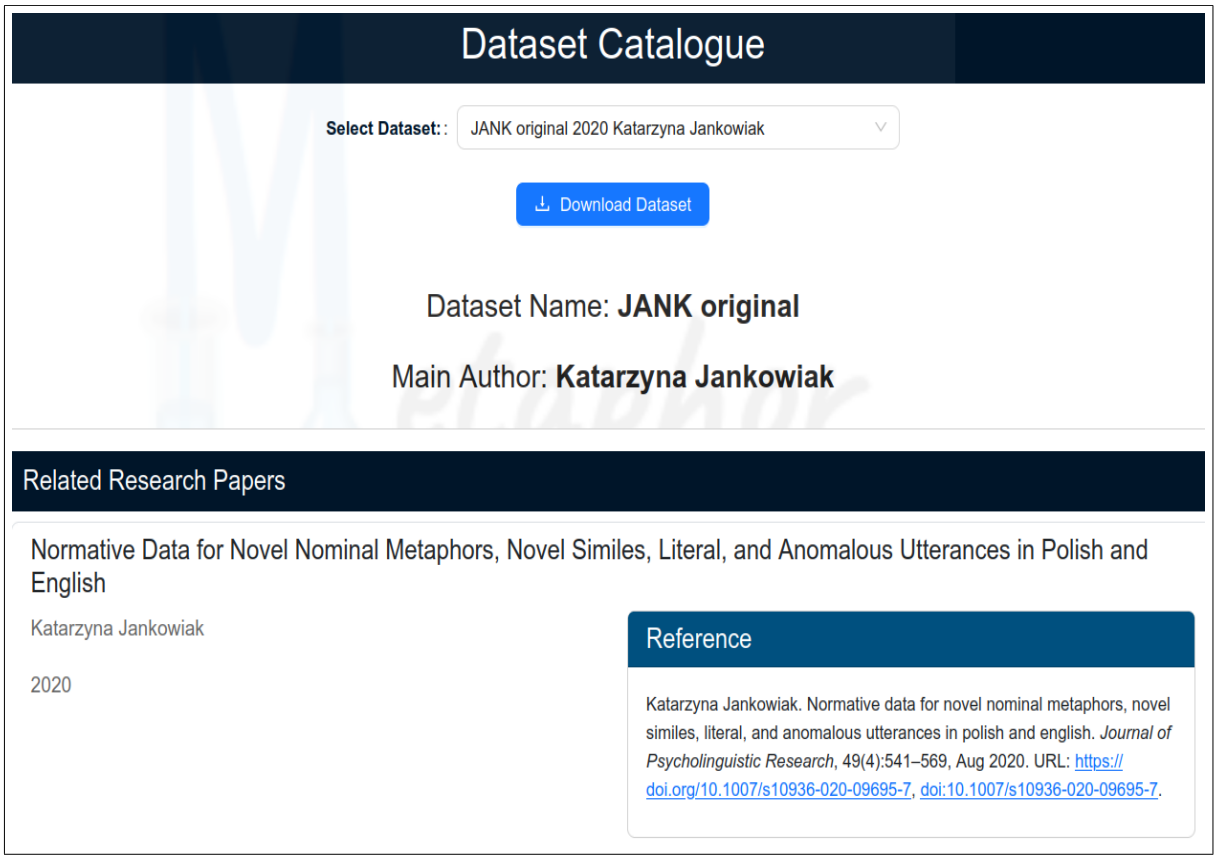

Figure 4: Screenshot of top of the the datasets information page in the catalog section of the website. The English dataset released with Jankowiak (2020) is presented as an example.

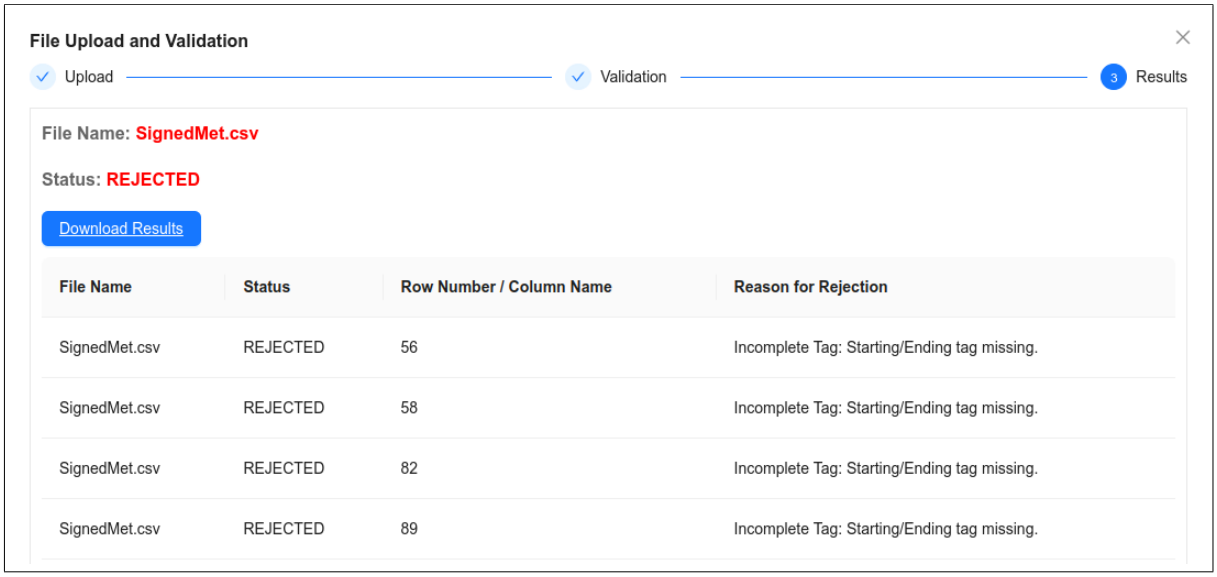

Figure 5: Screenshot of the file format check for a rejected file. The line the error occurs in the CSV file and the type of errors are specified.

Figure 6: Screenshot of the file format check after a CSV file is accepted for manual review.

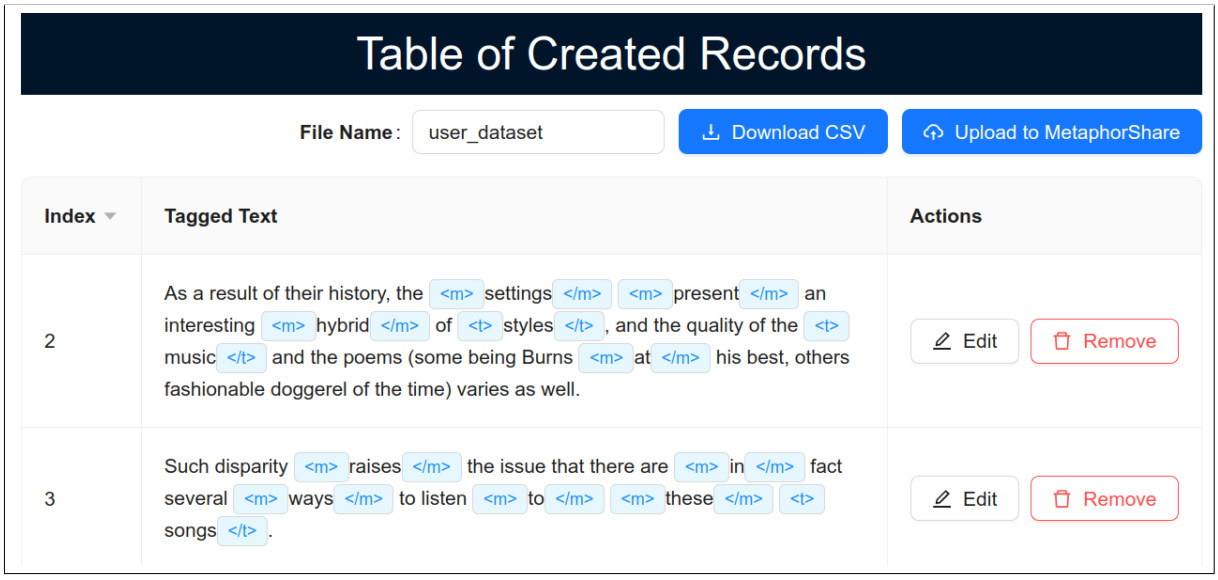

Figure 7: Screenshot showing dataset rows available for tagging or edition, as displayed in the annotation tool.