Developing Multilingual Multiword Expressions (MWEs) Resources: An Annotated Corpus of Fables following the PARSEME guidelines

Letícia Guedes Guimarães[1], Adriana Pagano[1], André V. Lopes Coneglian[1], Aline Villavicencio[2]

[1] Universidade Federal de Minas Gerais
[2]University of Exeter
{letigg, apagano, coneglian}@ufmg.br
a.villavicencio@exeter.ac.uk

Abstract. *This paper introduces a Brazilian Portuguese corpus of fables developed as part of a broader multilingual initiative. It comprises the narratives and their concluding moral statements (epimythia), annotated for morphosyntax and multiword expressions (MWEs) according to the Universal Dependencies (UD) and PARSEME guidelines, respectively. The analysis reveals that 53.59% of the sentences contain at least one MWE, with inherently reflexive verbs (IRVs) constituting the most frequent category. MWEs are also distributed differently: narratives contain a higher proportion of adverbial MWEs, whereas adpositional MWEs are frequent in epimythia. Future work will enrich the corpus through semantic annotation based on Uniform Meaning Representation (UMR).*

1. Introduction

Multiword expressions (MWEs) constitute a persistent challenge in Natural Language Processing (NLP) due to their idiosyncratic grammatical properties [Savary et al. 2018][Pasquer et al. 2020]. Their accurate identification is crucial for tasks such as parsing, machine translation, and semantic analysis, as failure to capture these expressions often leads to systematic errors in meaning interpretation.

Despite their importance in multilingual NLP resource development, corpora annotated with MWEs are limited across languages and text types [Ramisch et al. 2026]. Addressing this limitation requires text-type-diverse corpora; in this context, fables are particularly suitable for corpus construction because they are compact, self-contained discourse units that combine narrative discourse, dialogue, and reported speech with figurative and evaluative moral statements, thus moving from particular events to generalized meanings. This structure makes it possible to relate MWE use to different narrative stages and communicative functions. Moreover, fables circulate across languages and cultures, often in multiple adaptations, supporting multilingual and contrastive analysis.

This paper introduces an annotated corpus of Aesop's fables translated into Brazilian Portuguese as part of a larger multilingual corpus of fables translated into French, English, Spanish, Russian, Chinese, and Italian. The corpus comprises texts translated from Greek and annotated for morphosyntactic categories using the Universal Dependencies (UD) framework [Nivre et al. 2020] and for MWEs following the PARSEME guidelines [Savary et al. 2017] [Savary et al. 2026]. It is designed to support multilingual NLP analysis.

2. Related work

Annotated corpora enriched with linguistic information are essential for both training and evaluating Natural Language Processing (NLP) models [Ramisch et al. 2026]. A foundational initiative in this domain is the UD framework [Nivre et al. 2020], which provides standardized morphosyntactic annotation across languages, enabling consistent training and evaluation of multilingual parsers and treebanks. Building on UD, the PARSEME initiative [Savary et al. 2017] introduced guidelines for the annotation of MWEs, initially focusing on verbal MWEs and later extending to other categories, including nominal and adjectival expressions [Scholivet et al. 2026]. This effort resulted in gold-standard corpora that are widely used to evaluate MWE identification systems across multiple languages [Savary et al. 2026]. Details of the Brazilian Portuguese annotation process in Parseme versions 1.0 and 2.0, including challenging annotation cases, are provided in [Ramisch et al. 2018] and [Savary et al. 2026]; annotation has so far focused primarily on news reports, which highlights the need for annotated corpora covering other text types, which we addtress by annotating fables.

3. Methodology

For this multilingual corpus, the Brazilian Portuguese translations used are those published in *Esopo: fábulas completas* [Dezotti et al. 2013]. As the fables were available only in print, the first step consisted in manually transcribing them in a text editor. The texts were saved as plain-text (.txt) files and organized in two subsets: narratives and morals (epimythia). Overall, the corpus comprises 418 sentences and 8443 tokens excluding punctuation. The distribution of both subsets can be found in Table 1.

Table 1. Corpus statistics

Corpus	Sentences	Tokens	Types	Lemmas	Avg. sentence length
Narratives	355	7196	2152	1668	20.27
Epimythia	63	1247	507	438	19.79
Full corpus	418	8443	2427	1843	20.2

Subsequently, all .txt files were automatically parsed using UDPipe [Straka et al. 2016], trained on the Petrogold corpus [Souza et al. 2021]. The resulting annotated CONLLU files were then uploaded to the platform Arborator Grew [Guibon et al. 2020] and manually revised following the Brazilian Portuguese guidelines for UD annotation [Duran 2021, Duran 2022].

The revised files were uploaded to FLAT (FoLiA Linguistic Annotation Tool) [van Gompel 2024] for manual annotation of MWEs. Following the PARSEME framework [Savary et al. 2026], the annotation process included all MWE categories established by the project, which encompass verbal, nominal, adjectival, adverbial, and functional expressions[1]. Table 2 presents all MWE categories identified in the corpus, together with illustrative examples. The full set of categories, along with def-

[1]In compliance with PARSEME guidelines for Brazilian Portuguese [Savary et al. 2026], idiomatic verb-particle constructions (IVPCs) were not annotated. The guidelines report only one such construction in the language, namely “jogar fora” (lit. ‘to throw outside’; idiomatically, ‘to throw away’ or ‘to discard’), which was reclassified under a different label (VID) and the IVPC category excluded.

initions, identification tests, and examples, can be found in the PARSEME repository [PARSEME Initiative 2026].

Table 2. PARSEME categories adopted in this study

Type	Label	Name	Example and gloss
Verbal	IRV	Inherent reflexive verb	*se queixar* 'to complain'
	LVC.full	Light-verb construction (full)	*fazer uma visita* (lit. 'to make a visit'; 'to pay a visit')
	LVC.cause	Light-verb construction (causative)	*dar preferência* (lit. 'to give preference'; 'to prefer')
	VID	Verbal idiom	*correr riscos* (lit. 'to run risks'; 'to take risks')
	IAV	Inherently adpositional verb	*tratar de* (lit. 'to treat of'; 'to deal with')
	MVC	Multi-verb construction	*quer dizer* (lit. 'to want to say'; 'to mean')
Nominal	NID	Nominal idiom	*cara de pau* (lit. 'wooden face'; 'shameless person')
Adjectival / Adverbial	AdjID	Adjectival idiom	*cheia de si* (lit. 'full of herself'; 'conceited')
	AdvID	Adverbial idiom	*de novo* (lit. 'of new'; 'again')
Functional	AdpID	Adposition idiom	*a partir de* 'from; starting from'
	ConjID	Conjunction idiom	*uma vez que* 'since; given that'
	IntjID	Interjection idiom	*pois é* 'indeed; well, yes'

The annotation was carried out by the first author - a PhD candidate in Linguistics - and subsequently revised by the co-authors, senior researchers in Linguistics and Computational Linguistics. A blind procedure was not adopted due to the high level of expertise demanded by the annotation task. Disagreements were resolved through discussion guided by the PARSEME guidelines [Savary et al. 2026]. The resulting revised XML files were converted to XLSX format using a customized Python script and stored in a dedicated folder.

Quantitative analysis was conducted using two customized Python scripts. The first processed the CoNLL-U narrative and epimythia files into CSV format to extract structural metrics, including token, type, and lemma counts, as well as average sentence length (Table 1). The second processed the XLSX files to compute absolute and relative MWE frequencies, average MWEs per sentence, sentence-level MWE occurrence, and average MWE length. A lemmatization dictionary was used to map surface variants to canonical forms - for example, the forms "se pôs" (REFL put.PST.PFV.3SG, 'began') and "se punha" (REFL put.PST.IPFV.3SG, 'was beginning') were lemmatized as "pôr-se" (put.INF-REFL, 'to begin'). The resulting XLSX files were also used for exploratory qualitative analysis.

4. Results and Discussion

4.1. MWE density and distribution

The corpus comprises 322 MWE occurrences, with 245 distinct canonical forms, as established through the lemmatization procedure described in the Methodology; on average, each sentence contains 0.77 MWEs. Overall, 224 sentences contain at least one annotated MWE, corresponding to 53.59% of the dataset (Table 3). When narratives and epimythia are considered separately, the relative frequency of MWE occurrences remains similar in both, indicating a consistent presence of multiword phenomena across the corpus.

Table 3. Distribution of MWEs across the corpus

	MWE occurrences			Sentences with MWE		
	Total	**Distinct**	**Avg. MWE length**	**Total**	**Rel. freq. (%)**	**Avg. MWEs per sentence**
Narratives	268	206	2.35	189	53.24	0.75
Epimythia	54	47	2.33	35	55.56	0.86
Full corpus	322	245	2.34	224	53.59	0.77

The average length of MWEs in the corpus is 2.34 tokens, with the longest instances spanning 5 tokens: the adverbial idioms (AdvID) "A bem de a verdade" (lit. 'for the good of the truth'; idiomatically, 'truth be told') and "Em o que respeita a" (lit. 'in what respects to'; idiomatically,'regarding'). A total of 72% of occurrences span two tokens, a pattern that is heavily influenced by category distribution, as inherently reflexive verbs (IRVs) and light-verb constructions (LVCs) account for a considerable part of all MWE occurrences. IRVs are strictly formed by two lexicalized (fixed) items: a verb + a clitic (e.g., "tornar-se" ('to become'). As for LVCs, their structure is more flexible, since they can include a determiner or a preposition between the semantically bleached verb and the abstract predicative noun, causing them to span up to 4 tokens in Brazilian Portuguese. However, in our corpus, 70% of the LVC occurrences span only 2 tokens, e.g., "ter ideias" (lit. 'to possess ideas'; idiomatically, 'to come up with ideas').

4.2. MWE category distribution

IRVs account for the majority of MWE occurrences in the corpus, totaling 102 occurrences, with 73 distinct MWE lemmas. Overall, they account for 31.68% of all MWE instances. This MWE category is seconded by LVC.full (16.46%). The total number of occurrences for each categoryis presented in Table 4.

The distribution of MWE categories varies between the two fable sections. As the comparative Table 5 shows, while the narratives feature all twelve categories identified in the corpus, the epimythia lack any instances of AdjIDs, IAVs, or MVCs. It is worth noting, however, that these three categories are also scarce within the narrative data. Accounting for the scarcity of AdjIDs is beyond the scope of this paper, but the low frequency of IAVs and MVCs can be partially explained by both the annotation guidelines and linguistic factors.

In the case of IAVs, the primary cause likely stems from the annotation guidelines. This category covers verbs that require an adposition to exist or to hold a specific

Table 4. Distribution of MWE categories in the full corpus

	Occurrences			
Category	**Total**	**Rel. freq. (%)**	**Distinct**	**Rel. freq. (%)**
IRV	102	31.68	73	29.67
LVC.full	53	16.46	50	20.33
AdvID	48	14.91	35	14.23
AdpID	46	14.29	26	10.57
VID	30	9.32	27	11
IAV	12	3.73	7	2.85
ConjID	9	2.80	7	2.85
AdjID	8	2.48	8	3.25
NID	5	1.55	5	2.03
IntjID	5	1.55	4	1.63
LVC.cause	3	0.93	3	1.22
MVC	1	0.31	1	0.41

meaning, such as "tratar de" (lit.'to treat of'; idiomatically 'be about'). However, at the current stage of the project, this category remains experimental, and no language-specific guidelines have been established yet to systematically account for its occurrences. Consequently, the seven identified instances probably do not represent the full range of IAV constructions in the corpus, suggesting that their low frequency may partially reflect underannotation. As for MVCs, such as "querer dizer" (lit.'to want to say'; idiomatically, 'to mean'), there are only two occurrences of this category in the Brazilian Portuguese portion of the PARSEME corpus [Savary et al. 2026], suggesting that these constructions tend to be rare. The observed distribution in our corpus is consistent with this tendency.

Regarding the two remaining non-functional MWE categories, VIDs and NIDs, they account for only 10% of all occurrences, proving fairly rare in the corpus. Similarly to the case of AdjIDs, accounting for this scarcity is beyond the scope of this article; hence, a detailed investigation remains an upcoming step in our research.

4.3. Comparing narrative and epimythia: AdvIDs and AdpIDs

Although IRVs predominate across both fable sections, AdvIDs rank second in the narratives, whereas AdpIDs take second place in the epimythia (Table 5). This distribution directly reflects the distinct grammatical configurations characterizing these two functional sections [Neves 2014].

While narratives are primarily concerned with the unfolding of events, epimythia focus on expressing the moral lesson derived from them. Accordingly, narratives tend to favor adverbial idioms, which contribute to event sequencing, as in *em seguida* (lit. 'in sequence'; idiomatically, 'next' or 'then'); locate events in time and space, as in *tempos depois* (lit. 'times later'; idiomatically, 'some time later'); and express manner and situational circumstances, as in *em silêncio* (lit. 'in silence'; idiomatically, 'silently') and *de presente* (lit. 'of gift'; idiomatically, 'as a gift').

In contrast, epimythia show a stronger preference for adpositional idioms, which

encode relations of reference, as in *em relação a* (lit. 'in relation to'; idiomatically, 'with regard to'); opposition, as in *ao contrário de* (lit. 'to the contrary of'; idiomatically, 'in contrast to'); cause, as in *devido a* (lit. 'due to'; idiomatically, 'because of'); and purpose, as in *em busca de* (lit. 'in search of'; idiomatically, 'seeking'). These relations play a central role in the formulation of arguments and generalizations. This interpretation is further supported by the near absence of AdvIDs in epimythia, where the only occurrence is *às claras* (lit. 'at the clear [things]'; idiomatically, 'openly' or 'in plain sight'). In narratives, by contrast, AdpIDs primarily provide spatial, causal, and instrumental framing, as in *ao pé de* (lit. 'at the foot of'; idiomatically, 'next to'), *diante de* (lit. 'in front of'; idiomatically, 'faced with'), and *por meio de* (lit. 'by means of'; idiomatically, 'through').

Table 5. Distribution of MWE categories in narratives and epimythia

	Narratives			Epimythia			
Category	**Total**	**%**	**Rank**	**Total**	**%**	**Rank**	**Rank diff. (N−E)**
IRV	84	31.34	1	18	33.33	1	0
AdvID	47	17.54	2	1	1.85	9	-7
LVC.full	46	17.16	3	7	12.96	4	-1
AdpID	35	13.06	4	11	20.37	2	2
VID	21	7.84	5	9	16.67	3	2
IAV	12	4.48	6	0	0.00	—	—
AdjID	8	2.99	7	0	0.00	—	—
ConjID	7	2.61	8	2	3.70	5	3
NID	3	1.12	9	2	3.70	5	4
IntjID	3	1.12	9	2	3.70	5	4
LVC.cause	1	0.37	10	2	3.70	5	5
MVC	1	0.37	10	0	0.00	—	—

5. Conclusion and future work

This paper presented a manually annotated corpus of Aesop's fables translated into Brazilian Portuguese, developed as part of a broader multilingual initiative. The corpus was annotated for UD [Nivre et al. 2020] and for MWEs following the PARSEME guidelines [Savary et al. 2017][Savary et al. 2026], aiming to support multilingual and cross-text type NLP research.

The quantitative analysis revealed that 53.59% of the sentences contain at least one annotated MWE. IRVs account for the majority of instances across both narratives and epimythia. However, when comparing these two fable sections, adpositional MWEs are more frequent in the epimythia, whereas narratives favor adverbial MWEs. This reflects the distinct grammatical configurations of these two sections [Neves 2014].

Future work includes enriching the corpus with semantic annotation following the Universal Meaning Representation (UMR) framework [Chun and Xue 2024] and deepening the qualitative analysis to address open questions from the results and discussion section. The fully annotated dataset and the scripts used for data processing and analysis will be made publicly available in a GitHub repository under CC BY 4.0 and the MIT License, respectively.

6. Acknowledgments

The authors express their gratitude to the anonymous reviewers for their constructive feedback. Letícia Guedes acknowledges the financial support of the Minas Gerais State Agency for Research and Development (FAPEMIG). Adriana Pagano's research is funded by FAPEMIG and the National Council for Scientific and Technological Development (CNPq, grants 404722/2024-5; 313103/2021-6). André V. Lopes Coneglian holds a grant from CNPq (153972/2025-4). This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology) and in conducted under the auspices of the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILDIAR), funded by CNPq (grant 408490/2024-1).

7. Limitations

The relatively small size of the corpus restricts broader linguistic generalizations.

8. Ethics statement

The study does not involve human participants, personal data, or any type of sensitive information. All materials used consist solely of publicly available linguistic resources, including a book whose translated version was used with the explicit permission of its translator.

9. Generative AI usage

Generative AI tools were used in this work to generate Python code for data analysis and to assist with structuring sections and improving English language expression. All content was subsequently reviewed and validated by the authors. No content or interpretations were generated by AI.

References

Chun, J. and Xue, N. (2024). Uniform meaning representation parsing as a pipelined approach. In Ustalov, D., Gao, Y., Panchenko, A., Tutubalina, E., Nikishina, I., Ramesh, A., Sakhovskiy, A., Usbeck, R., Penn, G., and Valentino, M., editors, *Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing*, pages 40–52, Bangkok, Thailand. Association for Computational Linguistics.

Dezotti, M. C. C., Berliner, E., and Duarte, A. (2013). *Esopo: fábulas completas*. Cosac Naify, São Paulo, Brazil.

Duran, M. (2021). Manual de anotação de pos tags: orientações para anotação de etiquetas morfossintáticas em língua portuguesa, segundo as diretrizes da abordagem universal dependencies. Technical Report Relatório Técnico No. 434, Universidade de São Paulo, ICMC. Accessed: 2026-04-10.

Duran, M. (2022). Manual de anotação de relações de dependência: versão revisada e estendida. Technical Report Relatório Técnico No. 440, Universidade de São Paulo, ICMC. Accessed: 2026-04-30.

Guibon, G. et al. (2020). When collaborative treebank curation meets graph grammars. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 5291–5300.

Neves, M. H. M. (2014). A gramática pela fábula. ou: A fábula pela gramática. *Linguística*, 30(1):165–196.

Nivre, J. et al. (2020). Universal dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.

PARSEME Initiative (2026). Parseme corpora wiki: Citing parseme. `https://gitlab.com/parseme/corpora/-/wikis/home#citing-parseme`. Accessed: 2026-05-19.

Pasquer, C., Savary, A., Ramisch, C., and Antoine, J.-Y. (2020). Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, Barcelona, Spain. HAL Id: hal-03013636.

Ramisch, C. et al. (2018). A corpus study of verbal multiword expressions in brazilian portuguese. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018*, Lecture Notes in Artificial Intelligence, Cham, Switzerland. Springer International Publishing.

Ramisch, R., Ramisch, C., and Villavicencio, A. (2026). Expressões multipalavras: fogo de palha ou osso duro de roer? In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, volume 1, book chapter 5. BPLN, 4 edition.

Savary, A. et al. (2017). The parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions*, pages 31–47. Association for Computational Linguistics.

Savary, A. et al. (2018). Parseme multilingual corpus of verbal multiword expressions. In Markantonatou, S., Ramisch, C., Savary, A., and Vincze, V., editors, *Multiword Expressions at Length and in Depth: Extended Papers from the MWE 2017 Workshop*, pages 87–147. Language Science Press, Berlin.

Savary, A., Scholivet, M., Ramisch, C., Nakamura, T., Bilinski, E., Stymne, S., Giouli, V., Markantonatou, S., Pais, V., Mitrofan, M., Estève, L., Guillaume, B., Mititelu, V. B., Čibej, J., Hernández, R. D., Fendel, V., Gantar, P., Kanishcheva, O., Krstev, C., Liebeskind, C., Lobzhanidze, I., Marković, A. M., Nešpore-Bērzkalne, G., Pagano, A. S., Shamsfard, M., Stankovic, R., Tajalli, V., Tiberius, C., and Padhye, A. (2026). Parseme 2.0 multilingual corpus of multiword expressions. In Piperidis, S., Bel, N., van den Heuvel, H., Ide, N., Krek, S., and Toral, A., editors, *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4819–4834, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Scholivet, M., Savary, A., Ramisch, C., Bilinski, E., Nakamura, T., Mitrofan, M., and Pais, V. (2026). Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions. In Ojha, A. K., Mititelu, V. B., Constant, M., Stoyanova, I., Doğruöz, A. S., and Rademaker, A., editors, *Proceedings of the 22nd*

Workshop on Multiword Expressions (MWE 2026), pages 254–275, Rabat, Marocco. Association for Computational Linguistics.

Souza, E., Silveira, A., Cavalcanti, T., Castro, M., and Freitas, C. (2021). PetroGold – corpus padrão ouro para o dominio do petroleo. In Ruiz, E. E. S. and Torrent, T. T., editors, *Proceedings of the 13th Brazilian Symposium in Information and Human Language Technology*, pages 29–38, Porto Alegre, Brazil. Association for Computational Linguistics.

Straka, M., Hajič, J., and Straková, J. (2016). Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 4290–4297.

van Gompel, M. (2024). Folia linguistic annotation tool (version 0.11.5). KNAW Humanities Cluster CLST, Radboud University.