Journal of Open Humanities Data

Bearing a Bag-of-Tales: An Open Corpus of Annotated Folktales for Reproducible Research

RESEARCH PAPER

JOSHUA HAGEDORN
SÁNDOR DARÁNYI
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Motifs in folktales and myths have been identified and articulated by scholars, and the computational identification and discovery of such motifs is an area of ongoing research. Achieving this goal means meeting scientific requirements (that methods be comparable and replicable) and requirements for collaboration (that multi-disciplinary teams can reliably access data). To support those requirements, access to consistent reference datasets is needed. Unfortunately, these datasets are not openly available in a format that supports their use in data science. Here we report work in progress toward this goal, having converted the Ashliman Folktexts collection into a public dataset of annotated tale texts. The data can be accessed at doi.org/10.5281/zenodo.6575263.

CORRESPONDING AUTHOR:
Joshua Hagedorn
Independent researcher, Grand Rapids, MI, USA
josh.hagedorn@gmail.com

KEYWORDS:
annotated folktales; motifs; reproducible research; machine learning; version control

TO CITE THIS ARTICLE:
Hagedorn, J., & Darányi, S. (2022). Bearing a Bag-of-Tales: An Open Corpus of Annotated Folktales for Reproducible Research. *Journal of Open Humanities Data*, 8: 16, pp. 1–10. DOI: https://doi.org/10.5334/johd.78

(1) CONTEXT AND MOTIVATION

Ever since the concept of a motif was introduced some 200 years ago, the quest to identify elements of content above the word level has been a standard preoccupation in literary science (Frenzel, 1992; Seigneuret, 1988). There, a motif stands for a recurrent theme, whereas in musicology a motif is considered "the smallest structural unit possessing thematic identity" (White, 1976: 26–27). In the field of folktale research, Stith Thompson defined motifs as "the smallest element in a tale having a power to persist in tradition" (Thompson, 1977: 415).

The overlap between these definitions suggests that such higher-order content units exist as narrative building blocks, yet their automatic extraction by computational means has eluded folk narrative studies so far (Darányi & Lendvai, 2010). As we will argue below, folk narrative studies are not yet up to the task of a scalable pattern hunt. One reason for our scepticism is that in Thompson's Motif Index of Folk Literature (Thompson, 1951) alone over 45,000 motifs are listed on a global scale, but many more regional motif indexes exist whose material would doubtlessly inflate that number. If we want to apply machine learning for motif identification or discovery, first we need suitable datasets which enable research teams to replicate each other's results. Below we report work in progress in this direction, but also guard against any hubris in our promises regarding motif detection, with its first analytical results to be reported elsewhere. Ideally we would like to see an emerging motif annotation system that crowd-sourced expert folklorists could use, similar to Prodigy.[1]

The structure of this paper is as follows. In Section 1, we offer our motivation in context, bring examples of related research with converging trends, publicly available databases and datasets, and introduce the Ashliman Folktexts collection. Section 2 focuses on methodology, progressing from our motivation to support reproducible research in computational folkloristics toward dataset creation and repository access, including steps of data harvesting and cleaning, concluding with current limitations of use. Section 3 discusses features of the result, the Annotated Folktales (`aft`) corpus, with descriptive statistics. In Section 4, we briefly outline directions of future collection development to support folktale research.

(1.1) RELATED RESEARCH

As our pilot was not concerned with the structural analysis of folk narratives, this overview omits significant research results, such as those concerning the automatic detection of Proppian functions (Finlayson, 2016), or their use in ontology building (Declerck et al., 2017). Instead, our focus will be on precursory efforts to support motif detection using two standard tools, the Thompson Motif Index (TMI) (Thompson, 1951), and the Aarne-Thompson-Uther tale typology (ATU) (Uther, 2004). Important extensions to these, and to our current work, exist by Declerck and colleagues (Declerck & Schäfer, 2017; Declerck, Kostova, & Schäfer, 2017). As motifs and motifemes abound in myths as well, we admit the latter into our scope under the reasoning that "myth is a traditional tale with secondary, partial reference to something of collective importance" (Burkert, 1982: 23), considering the debate about the difference between myths and folktales as open (e.g. Kirk, 1970: 31–41; Burkert, 1982: 1–5).

(1.1.1) Converging trends

We consider finding characteristic patterns of semantic content by automatic means an open research problem. The relevant research question is this: if we were going to extract features from the descriptive text of the TMI, what kind of features could we build, and could these features also be identified in tale corpora?

The convergence of two major trends in computational folkloristics (Abello, Broadwell, & Tangherlini, 2012) will likely shape the results of the next decade. The first is a focus on the evolutionary aspect of motif and/or tale type distributions, either with regard to certain tale types (Bortolini et al., 2017; Karsdorp, 2016; Karsdorp & van den Bosch, 2013; da Silva & Tehrani, 2016; Tehrani, 2013), or to the geographical distribution of globally occurring narrative motifs (Thuillard, d'Huy, Berezkin, & Le Quellec, 2018), even inferring the presence of lost narratives (Kestemont et al, 2022). A genetic metaphor seems to inform some approaches, perhaps inspired

1 https://spacy.io/universe/project/prodigy (last accessed: 12 May 2022).

DOI: 10.5334/johd.78

by the modelling capacities inherent in Dawkins' meme theory (Dawkins, 1976); these compare tale types as motif sequences to 'narrative DNA' (Darányi, Wittek, & Forró, 2012; Meder et al., 2016; Murphy, 2015; Ofek, Darányi, & Rokach, 2013), or look at the evolution of narrative/story networks as a quasi-biological process based on the mutation and recombination of narrative elements (Karsdorp, 2016; Karsdorp & Fonteyn, 2019), extended even to the framework of cultural evolution via population genetics (Ross, Greenhill & Atkinson, 2013; Ross & Atkinson, 2015). Such methods resemble bioinformatic applications such as network motif identification (Qin & Gao, 2012), a problem analogous with ours. The context is that of *evolving semantics*, an emerging research area both in lexical semantic change (Armaselu et al., 2021) and digital preservation (Kontopoulos et al., 2016a; Kontopoulos et al., 2016b).

The second trend is to use probabilistic and/or multivariate statistical methods for the analysis of binary versus non-binary matrices of events over cases, where events can be index terms, motifs, motif sequences, and so on, and cases as an umbrella term stand for documents in general, such as abstracts describing narratives (Berezkin, 2015), or tale types (Uther, 2004), ultimately constituting text corpora or databases. On such collections, one can then experiment for instance with sub-corpus topic modelling (STM) by Latent Dirichlet Allocation (LDA) as a means of supervised passage exploration in partly unknown corpora (Tangherlini & Leonard, 2013).

The little one can say about the plethora of methods listed is that, regardless of the corpora, their regionality, and the analytical units whose distributions characterise the body of texts in question, they express similarity between items in terms of distance, with more similar items forming dense groups as the outcome of mass comparison. Cluster analysis (Thuillard et al., 2018), Principal Component Analysis (PCA) (Berezkin, 2015), Labelled Latent Dirichlet Allocation (L-LDA) (Karsdorp & van den Bosch, 2013), Support Vector Machines (SVM) (Nguyen et al., 2012; Meder et al., 2016), or deep learning by Recurrent Neural Networks (RNN) (Lô, de Boer, & van Aart, 2020), however, share the same nature of being static snapshots of collections. Hence there is a contradiction in principle in addressing text evolution, a dynamic phenomenon, through tools tailored to static measurements: the notion asks for vector fields instead of vector spaces (Darányi et al., 2016). The most promising recent direction seems to be the combination of word embeddings – increasingly condensed and geometrically located types of word meaning (Le & Mikolov, 2014; Mikolov et al., 2013; Reimers & Gurevych, 2019) – with deep learning: Pompeu (2019) reports successful application of a Hierarchical Attention Network (HAN) for the prediction of ATU categories on a multilingual database of folk texts.

As the computing of results for both trends discussed above require datasets, the next section briefly addresses their availability.

(1.1.2) Databases and datasets

Progress in computational folkloristics requires that results be replicable. To this end we sought open access datasets of ATU-annotated tales in English, but could not identify suitable candidates on GitHub,[2] Kaggle[3] or Google,[4] although websites with separate tale collections are available.[5] Neither could we find the big folklore data anticipated by Tangherlini & Leonard (2013) and Tangherlini (2016). Based on Meder (2010) and Ilyefalvi (2018), the largest databases seem to be the Dutch Folktale Database of the Meertens Institute, and the Danish Folklore Archive's Tang Kristensen Collection, the former in the magnitude of around 50,000 texts, the latter at around 34,000 texts (Tangherlini & Leonard, 2013). Other important databases exist (Berezkin, 2017), but are either beyond public access, or in their original languages only, or both. The notable exception is the Meertens Institute whose texts are in Dutch and Frisian plus a number of local dialects, but can be read in English translation as well.

Other researchers who have shared their data as supporting material for their articles include for instance Bortolini et al. (2017), da Silva & Tehrani, (2016), Tehrani (2013), and Tehrani,

2 https://github.com/awesomedata/awesome-public-datasets (last accessed: 4 May 2022).

3 https://www.kaggle.com/datasets (last accessed: 4 May 2022).

4 https://datasetsearch.research.google.com/ (last accessed: 4 May 2022).

5 https://fairytalez.com/authors-and-collections/ (last accessed: 1 June 2022).

Nguyen, & Roos (2016). Declerck et al. (2017) also report that a large amount of ATU data has recently been made available online by the Multilingual Folk Tale Database (MFTD),[6] which also offers annotation facilities for tales in multilingual versions. We found only a single recent study (Lô, de Boer, & van Aart, 2020) which published a corresponding tale corpus to promote reproducibility, albeit without ATU type labels.[7]

Among the ATU-annotated tale collections publicly available on the internet, the most promising candidate was Ashliman's Folktexts collection. The process of the conversion of this collection to the desired format will be described below.

(1.1.3) The Ashliman Folktexts collection

The Folktexts site[8] has been populated and maintained since 1996 by D.L. Ashliman, who kindly agreed to donate his collection to the interested research communities.[9] While other sites may sport a more lavish design, this one is the largest and most extensively annotated. It serves as a respected scholarly resource for folklorists, with a large and curated set of tale texts. Whereas our dataset contains only tales from pages with clear ATU annotations (214 pages), the total content of the website is much larger (370 pages), containing various creation myths, stories of changelings, Faust legends, and Christiansen's tale types (Christiansen, 1992). However, it is the ATU annotation that makes this corpus particularly valuable as a potential training dataset for classification methods.

Despite the richness of this resource, it has not frequently been used in folklore research as a larger corpus. Some previous studies reference it, yet these often only include a smaller portion of the entire set of texts (Reiter, Frank, & Hellwig, 2014). To the best of our knowledge, none of the published studies provided open access to the data.

(2) METHOD

(2.1) SUPPORT FOR REPRODUCIBILITY IN FOLKLORE STUDIES

Reproducibility is a defining characteristic of science, yet a wide gamut of scientific fields has been plagued by a 'replicability crisis': a situation where trusted research findings have been impossible to reproduce (Goodman, Fanelli, & Ioannidis, 2016; Pasquier et al., 2017). While the problem has come to the fore in the health and social sciences, it has been acknowledged in disciplines as broad as archaeology (Marwick, 2017), public health (Harris et al., 2018), biology (Kühne & Liehr, 2009), and economics (McCullough, 2009).

Reproducible research entails that study results be accompanied by (Gandrud, 2015):

1. a detailed description of the methods used to obtain and operate on the data;
2. the full dataset(s) used in the study;
3. the full code used to transform the data and compute the results.

(2.1.1) Guiding principles

The following features guided our selection of tools and format for the code and data:

- *Open data*: In order to use tale data consistently, it must be made freely and openly available to anyone. The dataset is therefore distributed under a Creative Commons Attribution-ShareAlike 4.0 International license (CC BY-SA 4.0).[10]
- *Extensible data*: The dataset can be added to or modified, in order to develop a more complete repository of tales. This can be done by submitting pull requests to the project's GitHub repository.
- *Open code*: Any user is allowed to view and run the code that produces the dataset, as well as downstream analyses which use the dataset. This allows for inspection,

6 https://www.mftd.org (last accessed: 4 May 2022).

7 https://github.com/GossaLo/afr-neural-folktales/ (last accessed: 4 May 2022).

8 http://www.pitt.edu/~dash/folktexts.html (last accessed: 1 June 2022).

9 Personal communication per email (11 February 2021).

10 https://creativecommons.org/licenses/by-sa/4.0/ (last accessed: 4 May 2022).

refinement, and reasoning about the effects of transformation and statistical modelling on the data.

- *Common form*: We have chosen to use the "tidy" dataframe as the structure of the dataset, in which (a) each variable forms a column, (b) each observation forms a row, and (c) a single type of observational unit forms the dataframe (Wickham, 2014).
- *Common tools*: The data must also be structured in a way that allows for ease of use with the standard tools of the trade of data science, such as R or Python.
- *Modifiable form*: The structure must allow for reshaping the data into sparse matrices, nested structures, and graph-based structures as dictated by the needs of a given text analysis, while starting from a common source dataset (that is, the `aft`).

(2.1.2) Accessing and growing the corpus

Snapshot versions of the `aft` corpus will be cached on Zenodo with development and collaboration ongoing in the `trilogy` GitHub repository, where a vignette provides information on how to access, use, and augment the dataset.[11] Whereas long-term sustainability to curate the result will require academic resources, as a next step it would be logical to create temporary merger options with other multilingual tale collections, such as the MFTD, or the ones analyzed by Tehrani (2013) or Karsdorp and Fonteyn (2019), for analytical studies using for instance multilingual word embeddings.

The open-source *Git* functionality allows motifs, tale types and annotated tales to be added over time, and for the corpus to serve as a communal resource. We welcome inquiries and suggestions about how best to manage this resource as a "commons" (Vollan & Ostrom, 2010).

(2.2) DATA HARVESTING AND CLEANING

(2.2.1) Steps

Web-scraping of the Folktexts site[12] was completed using the `R` statistical programming language. The following high-level summary is provided to allow for an understanding of the methods used and their limitations:

1. Obtain URLs and associated label text for all 'child' pages of the main website to create a dataframe of page names and URLs, removing links to external websites.
2. Retain all URLs with the pattern "`type…`", which denote pages containing tales which belong to an ATU type, and recode links which do not follow this form, such as the page for *Animal Brides and Animal Bridegrooms* which was recoded as belonging to ATU type 402.
3. Extract the ATU type ID from the URL for each page, resulting in a dataframe listing 214 webpages, each associated with a tale type and containing the page name, page URL, and associated ATU ID.
4. Loop through each webpage identified in the dataframe above and extract the text, using the following steps: (a) extract HTML nodes from the page, creating a dataframe using the text, name and attribute elements of the nodes; (b) remove superfluous text other than tale texts, titles, and other associated metadata (e.g. source documents, notes); (c) using a fuzzy-joining method to align missing body text with the well-formatted HTML.[13]
5. Take the resulting dataframe and apply the following steps: (a) select the longest `text`, choosing between the tagged HTML version and the version extracted from the `body`; (b) select available metadata; (c) remove irrelevant entries using regular expressions; (d) create unique tale titles where these were duplicated across multiple variants of tales; (e) clean tale text data (e.g. removing remnant HTML tags, replacing internal double quotes with single quotes).
6. Add manually extracted tales into a consistent format for web pages which generated errors during web scraping (ATU IDs 1696, 2, 545B, 57, 675, 75, 779J*, 676). Other than this final step, all steps were fully automatic.

11 https://github.com/j-hagedorn/trilogy/blob/master/docs/vignettes/getting_started.md (last accessed: 4 May 2022).

12 The site was harvested on 10 March 2021.

13 Using the *Jaro-Winkler* method, with maximum distance for a match set to 1 (Winkler, 1990; van der Loo, 2014).

(2.2.2) Limitations

Web-scraping is an inherently messy exercise, as the data contained in web pages are often not formatted with the intent of being analysed. While the output has been reviewed at a cursory level, we anticipate that greater use of the dataset will result in the need for additional cleaning and processing.

The `provenance` field does not meet the definition of 'tidy' outlined above, since multiple types of descriptors (e.g. *country*, *region*, *tale collection*) are stored in a single column. While additional cleaning may be able to distinguish some of these, we have chosen to leave it as entered in the original to avoid losing potentially valuable detail.

The final limitation is purposefully adopted for the sake of downstream analyses. We have included only tales which were annotated with a single tale type, despite the existence of some tales which can be characterized by multiple types. This decision was made in order to avoid repeating texts or using data structures which are tool specific.[14]

(3) RESULTS AND DISCUSSION

(3.1.) FEATURES OF THE ANNOTATED FOLKTALES (`aft`) DATASET

(3.1.1) Data dictionary

The `aft` (henceforth standing for *Annotated Folktales* to allow for the future inclusion of other resources) dataframe contains 1518 rows, each corresponding to a single tale. Its eight columns are described briefly below:

- `atu_id`: The ATU tale type identifier which classifies the tale.
- `tale_title`: The title of the tale.
- `provenance`: The person, place or tradition from which the tale came. In Ashliman's collection, this refers variously to the person recording the tales (e.g. Giambattista Basile), the country or region from which the version of the tale came (e.g. North Africa), or the larger collection of tales in which the tale is found (e.g. the Kathasaritsagara).
- `notes`: Additional notes related to the tale.
- `source`: The bibliographic citation for the original published source of the tale.
- `text`: The full text of the tale identified in `tale_title`.
- `data_source`: The source of the annotated tales. At the time of this writing, the source of all tales is Ashliman's Folktexts, but this is intended to change as the dataset grows.
- `date_obtained`: The date on which the dataset identified as a `data_source` was last downloaded and compiled.

Table 1 below shows the initial characters of fields from the first six rows of the dataset, in order to illustrate its appearance:

Table 1 Example output of the dataset.

atu_id	tale_title	provenance	source	text
910B	The Highlander Takes...	Scotland	Cuthbert Bede	In one of the glens of...
910B	The Prince Who Acquired	India	Cecil Henry Bompas	There was once a raja ...
910B	The Three Admonitions	Italy	Thomas Frederick Crane	A man once left his co...
910B	The Three Advices	Ireland	T. Crofton Croker	The stories current am...
910B	The Three Advices Which...	Ireland	Patrick Kennedy	The name of the young ...
1430	Buttermilk Jack		Thomas Hughes	Oh mother, my buttermilk

(3.1.2) Descriptive statistics

The 1518 tales in the dataset average 979.1 tokens in length, though the individual texts vary with a minimum of 10 tokens and a maximum of 12,406 (Table 2).

The histogram below (see Figure 1) shows the distribution of tale lengths for all tales in the corpus[15]:

14 The list structure is specific to R, and different in Python. Texts with multiple IDs would result in a nested list so we would stop being data-tool agnostic.

15 Excludes six tales with greater than 6,000 tokens, to increase visibility.

MEASURE	VALUE
Number of tales	1518
Number of tale types	182
Mean tokens per tale	979.1
Median tokens per tale	642
Minimum tokens per tale	10
Maximum tokens per tale	12,406
Mean sentences per tale	45.7
Median sentences per tale	31

Table 2 Summary statistics of the AFT dataset.

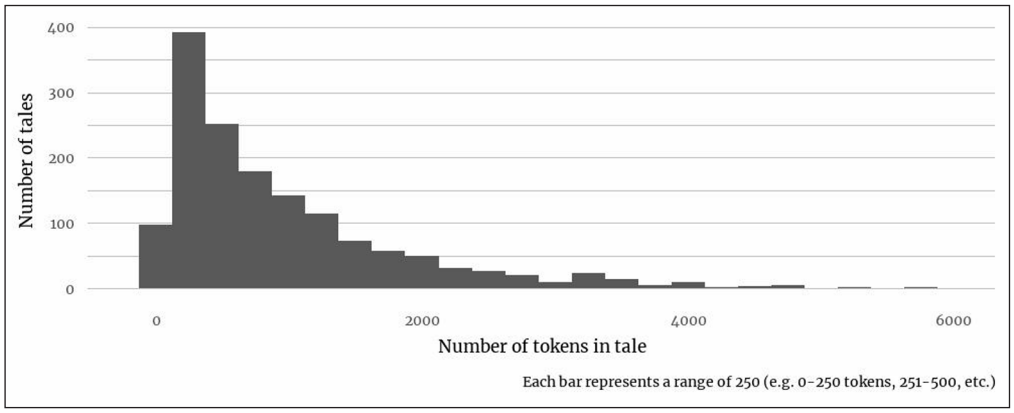

Figure 1 Distribution of tale lengths.

The tales compiled in the `aft` data are annotated by ATU tale type, and represent 182 distinct types. There are on average 8.3 tales in each tale type, with a range of one to 31. The tale types with the largest representative group of tales in the corpus are shown in Table 3 below:

ATU ID	TALE NAME	N OF TALES
275	The Race between Two Animals (previously The Race of the Fox and the Crayfish)	31
777	The Wandering Jew	30
1645	The Treasure at Home	26
510B	Peau d'Asne (previously The Dress of Gold, of Silver, and of Stars [Cap o Rushes])	26
500	The Name of the Supernatural Helper	23
510A	Cinderella	21
700	Thumbling (previously Tom Thumb)	21
155	The Ungrateful Snake Returned to Captivity	20
545B	Puss in Boots	20
980	The Ungrateful Son (previously Ungrateful Son Reproved by Naive Actions of Own Son)	20

Table 3 Ten tale types with the largest number of representative tales.

(4) IMPLICATIONS/APPLICATIONS

Under a Creative Commons license, we published on Zenodo and GitHub an open-access, ATU-annotated dataset of 1518 tales for motif detection by machine learning. This dataset resulted from the conversion of the Ashliman Folktexts collection, and is hoped to become the core of an expanding corpus to support reproducible research in computational folkloristics. As a next step we plan to integrate information from the TMI and the ATU, to be applied in trawling (Tangherlini & Leonard, 2013) for motifs by deep learning.

ACKNOWLEDGEMENTS

The authors are grateful to Professor D.L. Ashliman (University of Pittsburgh) for his permission to turn his annotated collection into a publicly available dataset. We also thank three anonymous reviewers for helpful comments on the manuscript.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Joshua Hagedorn: conceptualisation, methodology, data curation, software, validation, writing – original draft, writing – review & editing.

Sándor Daránÿi: conceptualisation, methodology, resources, supervision, project administration, writing – original draft, writing – review & editing.

AUTHOR AFFILIATIONS

Joshua Hagedorn orcid.org/0000-0001-8026-7562
Independent researcher, Grand Rapids, MI, US

Sándor Daránÿi orcid.org/0000-0002-1542-934X
Swedish School of Library and Information Science, University of Borås, Borås, SE

REFERENCES

Abello, J., Broadwell, P., & **Tangherlini, T. R.** (2012). Computational folkloristics. *Communications of the ACM*, *55*(7), 60–70. DOI: https://doi.org/10.1145/2209249.2209267

Armaselu, F., Apostol, E.-S., Khan, F., Liebeskind, C., McGillivray, B., Truica, C.-O., Utka, A., Oleškevičienė, G. V., & **van Erp, M.** (2022). LL(O)D and NLP perspectives on semantic change for Humanities research. *Semantic Web Journal*, accepted for publication. http://www.semantic-web-journal.net/system/files/swj2848.pdf. DOI: https://doi.org/10.3233/SW-222848

Berezkin, Y. (2015). Spread of folklore motifs as a proxy for information exchange: Contact zones and borderlines in Eurasia. *Trames*, *19*(1), 3–14. DOI: https://doi.org/10.3176/tr.2015.1.01

Berezkin, Y. (2017). Peopling of the New World from data on distributions of folklore motifs. In R. Kenna, M. MacCarron, & P. MacCarron (Eds.), *Maths meets myths: Quantitative approaches to ancient narratives* (pp. 71–89). Cham, Switzerland: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-39445-9_5

Bortolini, E., Pagani, L., Crema, E. R., Sarno, S., Barbieri, C., Boattini, A., Sazzini, M., da Silva, S. G., Martini, G., Metspalu, M., Pettener, D., Luiselli, D., & **Tehrani, J. J.** (2017). Inferring patterns of folktale diffusion using genomic data. *Proceedings of the National Academy of Sciences*, *114*(34), 9140–9145. DOI: https://doi.org/10.1073/pnas.1614395114

Burkert, W. (1982). *Structure and history in Greek mythology and ritual*. Berkeley, CA: University of California Press.

Christiansen, R. (1992). *The migratory legends: A proposed list of types with a systematic catalogue of the Norwegian variants*. Helsinki: Suomalainen Tiedeakatemia, Academia Scientiarum Fennica.

Daránÿi, S., & **Lendvai, P.** (Eds.). (2010). *Proceedings of the first AMICUS workshop, October 21, 2010, Vienna, Austria*. Szeged, Hungary: University of Szeged, Department of Library and Human Information Science.

Daránÿi, S., Wittek, P., & **Forró, L.** (2012). Toward sequencing 'Narrative DNA': Tale types, motif strings and memetic pathways. *Proceedings of CMN-12, 3rd workshop on Computational models of narrative in conjunction with the 8th Language resources and evaluation conference*, 2–10. http://narrative.csail.mit.edu/cmn12/proceedings.pdf

Daránÿi, S., Wittek, P., Konstantinidis, K., Papadopoulos, S., & **Kontopoulos, E.** (2016). A physical metaphor to study semantic drift. To appear in *Proceedings of SuCCESS-16, 1st international workshop on Semantic change & evolving semantics*. DOI: https://doi.org/10.48550/arXiv.1608.01298

da Silva, S. G., & **Tehrani, J. J.** (2016). Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. *Royal Society Open Science*, *3*(1), 1–11. DOI: https://doi.org/10.1098/rsos.150645

Dawkins, R. (1976). *The selfish gene*. Oxford, UK: Oxford University Press.

Declerck, T., Aman, A., Banzer, M., Macháček, D., Schäfer, L., & **Skachkova, N.** (2017). Multilingual ontologies for the representation and processing of folktales. *Proceedings of the First workshop on Language technology for Digital Humanities in Central and (South-)Eastern Europe*, 20–23. DOI: https://doi.org/10.26615/978-954-452-046-5_003

Declerck, T., Kostova, A., & **Schäfer, L.** (2017). Towards a linked data access to folktales classified by Thompson's motifs and Aarne-Thompson-Uther's types. *Proceedings of Digital Humanities 2017*. https://www.dfki.de/fileadmin/user_upload/import/9028_Dh2017_LOD_TMI-ATU_final.pdf

Declerck, T., & **Schäfer, L.** (2017). Porting past classification schemes for narratives to a linked data framework. *Proceedings of DATeCH2017*. DOI: https://doi.org/10.1145/3078081.3078105

Finlayson, A. M. (2016). Inferring Propp's functions from semantically annotated text. *Journal of American Folklore*, 55–77. DOI: https://doi.org/10.5406/jamerfolk.129.511.0055

Frenzel, E. (1992). *Stoffe der Weltliteratur: Ein Lexikon dichtungsgeschichtlicher Längsschnitte* (8. überarb. u. erweit. Aufl.). Stuttgart: Kröner.

Gandrud, C. (2015). *Reproducible research with R and RStudio* (2nd ed.). Chapman and Hall/CRC. DOI: https://doi.org/10.1201/9781315382548

Goodman, S. N., Fanelli, D., & **Ioannidis, J. P. A.** (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 341ps12. DOI: https://doi.org/10.1126/scitranslmed.aaf5027

Harris, J. K., Johnson, K. J., Carothers, B. J., Combs, T. B., Luke, D. A., & **Wang, X.** (2018). Use of reproducible research practices in public health: A survey of public health analysts. *PLoS ONE*, *13*(9). DOI: https://doi.org/10.1371/journal.pone.0202447

Ilyefalvi, E. (2018). The theoretical, methodological and technical issues of digital folklore databases and computational folkloristics. *Acta Ethnographica Hungarica*, *63*(1), 209–258. DOI: https://doi.org/10.1556/022.2018.63.1.11

Karsdorp, F. (2016). *Retelling stories: A computational-evolutionary perspective* [PhD thesis]. Nijmegen: Radboud Universiteit.

Karsdorp, F., & **Fonteyn, L.** (2019). Cultural entrenchment of folktales is encoded in language. *Palgrave Communications*, *5*, 25. DOI: https://doi.org/10.1057/s41599-019-0234-9

Karsdorp, F., & **van den Bosch, A.** (2013). Identifying motifs in folktales using topic models. In *Proceedings of the 22 Annual Belgian-Dutch Conference on Machine Learning* (pp. 41–49).

Kestemont, M., Karsdorp, F., De Bruijn, E., Driscoll, M., Kapitan, K. A., Ó Macháin, P., Sawyer, D., Sleiderink, R., & **Chao, A.** (2022). Forgotten books: The application of unseen species models to the survival of culture. *Science*, *375*(6582), 765–769. DOI: https://doi.org/10.1126/science.abl7655

Kirk, G. S. (1970). *Myth: Its meaning and functions in ancient and other cultures*. Berkeley, CA: University of California Press. DOI: https://doi.org/10.1525/9780520342378

Kontopoulos, E., Darányi, S., Wittek, P., Konstantinidis, K., Riga, M., Mitzias, P., Stavropoulos, T., Andreadis, S., Maronidis, A., Karakostas, A., & **others.** (2016a). *Deliverable 4.5: Context-aware content interpretation*. PERICLES project.

Kontopoulos, E., Riga, M., Mitzias, P., Andreadis, S., Stavropoulos, T., Konstantinidis, K., Maronidis, A., Karakostas, A., Tachos, S., Kaltsa, V., & **others.** (2016b). *Pericles deliverable 4.4: Modelling contextualised semantics*. PERICLES project.

Kühne, M., & **Liehr, A.** (2009). Improving the traditional information management in natural sciences. *Data Science Journal*, *8*, 18–26. DOI: https://doi.org/10.2481/dsj.8.18

Le, Q. V., & **Mikolov, T.** (2014). Distributed representations of sentences and documents. *Proceedings of the 31st International conference on Machine Learning*, PMLR *32*(2), 1188–1196. DOI: https://doi.org/10.48550/arXiv.1405.4053

Lô, G., de Boer, V., & **van Aart, C. J.** (2020). Exploring West African folk narrative texts using Machine Learning. *Information*, *11*(5), 236. DOI: https://doi.org/10.3390/info11050236

Marwick, B. (2017). Computational reproducibility in archaeological research: Basic principles and a case study of their implementation. *Journal of Archaeological Method and Theory*, *24*, 424–450. DOI: https://doi.org/10.1007/s10816-015-9272-9

McCullough, B. D. (2009). Open Access Economics journals and the market for reproducible economic research. *Economic Analysis and Policy*, *39*(1), 117–126. DOI: https://doi.org/10.1016/S0313-5926(09)50047-1

Meder, T. (2010). From a Dutch folktale database towards an international folktale database. *Fabula*, *51*(1-2), 6–22.

Meder, T., Karsdorp, F., Nguyen, D.-P., Theune, M., Trieschnigg, R. B., & **Muiser, I.** (2016). Automatic enrichment and classification of folktales in the Dutch Folktale Database. *The Journal of American Folklore*, *129*(511), 78–96. DOI: https://doi.org/10.5406/jamerfolk.129.511.0078

Mikolov, T., Chen, K., Corrado, G., & **Dean, J.** (2013). Efficient estimation of word representations in vector space. DOI: https://doi.org/10.48550/arXiv.1301.3781

Murphy, T. P. (2015). *From fairy tale to film screenplay: Working with plot genotypes*. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan UK. DOI: https://doi.org/10.1057/9781137552037

Nguyen, D., Trieschnigg, D., Meder, T., & **Theune, M.** (2012). Automatic classification of folk narrative genres. In J. Jancsary (Ed.), *Proceedings of KONVENS 2012* (pp. 378–382).

Ofek, N., Darányi, S., & **Rokach, L.** (2013). Linking motif sequences with tale types by Machine Learning. In M. A. Finlayson, B. Fisseni, B. Löwe, & J. C. Meister (Eds.), *2013 Workshop on Computational Models of Narrative, 32,* 166–182. DOI: https://doi.org/10.4230/OASIcs.CMN.2013.166

Pasquier, T., Lau, M. K., Trisovic, A., Boose, E. R., Couturier, B., Crosas, M., Ellison, A. M., Gibson, V., Jones, C. R., & **Seltzer, M.** (2017). If these data could talk. *Scientific Data*, 4, 170114. DOI: https://doi.org/10.1038/sdata.2017.114

Pompeu, D. (2019). *Interpretable Deep Learning methods for classifying folktales according to the Aarne-Thompson-Uther scheme*. Master's Thesis. Lisboa: Instituto Superior Técnico, Universidade de Lisboa.

Qin, G., & **Gao, L.** (2012). An algorithm for network motif discovery in biological networks. *IJDMB*, 1–16. DOI: https://doi.org/10.1504/IJDMB.2012.045533

Reimers, N., & **Gurevych, I.** (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. DOI: https://doi.org/10.18653/v1/D19-1410

Reiter, N., Frank, A., & **Hellwig, O.** (2014). An NLP-based cross-document approach to narrative structure discovery. *Literary and Linguistic Computing*, *29*(4), 583–605. DOI: https://doi.org/10.1093/llc/fqu055

Ross, R. M., Greenhill, S. J., & **Atkinson, Q. D.** (2013). Population structure and cultural geography of a folktale in Europe. *Proc R Soc B,* 280: 20123065. DOI: https://doi.org/10.1098/rspb.2012.3065

Ross, R. M., & **Atkinson, Q. D.** (2015). Folktale transmission in the Arctic provides evidence for high bandwidth social learning among hunter-gatherer groups. *Evolution and Human Behavior*, *37*(1), 47-53. DOI: https://doi.org/10.1016/j.evolhumbehav.2015.08.001

Seigneuret, J. C. (1988). *Dictionary of literary themes and motifs*. New York, NY: Greenwood Press.

Tangherlini, T. R. (2016). Big folklore: A special issue on computational folkloristics. *The Journal of American Folklore*, *129*(511), 5–13. DOI: https://doi.org/10.5406/jamerfolk.129.511.0005

Tangherlini, T. R., & **Leonard, P.** (2013). Trawling in the sea of the Great Unread: Sub-corpus topic modeling and Humanities research. *Poetics*, *41*(6), 725–749. DOI: https://doi.org/10.1016/j.poetic.2013.08.002

Tehrani, J. J. (2013). The phylogeny of Little Red Riding Hood. *PLoS ONE*, *8*(11), e78871. DOI: https://doi.org/10.1371/journal.pone.0078871

Tehrani, J. J., Nguyen, Q., & **Roos, T.** (2016). Oral fairy tale or literary fake? Investigating the origins of *Little Red Riding Hood* using phylogenetic network analysis. *Digital Scholarship in the Humanities.*, *31*(3), 611–636. DOI: https://doi.org/10.1093/llc/fqv016

Thompson, S. (1951). *Motif-index of folk-literature: A classification of narrative elements in folktales, ballads, myths, fables, mediaeval romances, example, fabliaux, jest-books and local legends* (rev. [2nd] ed.). Copenhagen: Rosenkilde.

Thompson, S. (1977). *The Folktale*. Berkeley, CA: University of California Press.

Thuillard, M., d'Huy, J., Berezkin, Y., & **Le Quellec, J.-L.** (2018). A large-scale study of world myths. *Trames*, *22*(4), 407–424. DOI: https://doi.org/10.3176/tr.2018.4.05

Uther, H.-J. (2004). *The types of international folktales: A classification and bibliography, based on the system of Antti Aarne and Stith Thompson*. Helsinki: Suomalainen Tiedeakatemia, Academia Scientiarum Fennica.

van der Loo, M. P. J. (2014). The stringdist package for approximate string matching. *The R Journal*, *6*(1), 111–122. DOI: https://doi.org/10.32614/RJ-2014-011

Vollan, B., & **Ostrom, E.** (2010). Cooperation and the Commons. *Science*, *330*(6006), 923–924. DOI: https://doi.org/10.1126/science.1198349

White, J. D. (1976). *The analysis of music*. New York, NY: Prentice-Hall.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(1), 1–23. DOI: https://doi.org/10.18637/jss.v059.i10

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey research methods*, 354–359. https://eric.ed.gov/?id=ED325505

TO CITE THIS ARTICLE:
Hagedorn, J., & Darányi, S. (2022). Bearing a Bag-of-Tales: An Open Corpus of Annotated Folktales for Reproducible Research. *Journal of Open Humanities Data*, 8: 16, pp. 1–10. DOI: https://doi.org/10.5334/johd.78

Published: 24 June 2022

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.