Noname manuscript No.
(will be inserted by the editor)

Bearing a bag-of-tales:

An open corpus of annotated folktales for reproducible research

Joshua Hagedorn · Sándor Darányi ·

Received: date / Accepted: date

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. The minimal reference datasets for motif identification include: known motifs (e.g. the Thompson Motif Index), known sequences of motifs occurring in tales (e.g. the Aarne-Thompson-Uther tale typology), and actual tale texts annotated with these motif sequences. Here we report work in progress toward this goal, having converted the Ashliman Folktexts collection into a public dataset of annotated tale texts.

Keywords annotated folktales · mythology · motif · reproducible research · machine learning · version control ·

Grants or other notes about the article that should go on the front page should be placed here. General acknowledgments should be placed at the end of the article.

Joshua Hagedorn
Independent researcher, Grand Rapids MI, USA
E-mail: josh.hagedorn@gmail.com

Sándor Darányi
Swedish School of Library and Information Science, University of Borås
E-mail: sandor.daranyi.hb.se@gmail.com

1 Introduction

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 [17], [43]. There, a motif stands for a recurrent theme, whereas in musicology a motif is considered "the smallest structural unit possessing thematic identity" [54]. In the vein of folktale research, Stith Thompson defined motifs as "the smallest element in a tale having a power to persist in tradition" [50].

The sufficient overlap between these definitions suggests that such higher-order content units exist as narrative building blocks in a generic sense, yet their automatic extraction by computational means has eluded folk narrative studies so far [5]. Despite the suggestion that topics identified by Labeled Latent Dirichlet Allocation (L-LDA) had an analogous function with motifs in a database of Dutch and Frisian folktales [22], we consider finding characteristic patterns of semantic content an open research problem. One reason for our skepticism is that in Thompson's Motif Index of Folk Literature [49] alone, over 45000 motifs are listed on a global scale, but many more e.g. regional motif indexes exist whose material would doubtlessly inflate that number. As we will argue below, digital humanities (DH) in general, and folk narrative studies in particular, are not up to the task of a scalable pattern hunt yet. If we want to apply machine learning for motif identification and extraction, we need suitable datasets which enable research teams to replicate each other's results. Below we report an initial step in this direction.

The structure of this paper is as follows. In Section 2, we bring examples of related research. In Section 3, Ashliman's Folktexts tale collection is introduced. In Section 4 we explain our motivation to support reproducible research in computational folkloristics, with Section 5 offering details of data harvesting and cleaning. Section 6 brings details about the new annotated dataset for machine learning, while in Section 7 we add our conclusions and plans for future research.

2 Related research

As our pilot was not concerned with the structural analysis of folk narratives, we left out from this brief overview significant research results concerning e.g. the automatic detection of Proppian functions [16], or their use in ontology building [12]. Instead, our focus will be on precursory efforts to automatic motif detection using two standard tools, the Thompson Motif Index (TMI) [49], and the Aarne-Thompson-Uther tale typology (ATU) [52]. Important extensions to these, and to our current work, exist e.g. by [14] and [13] .

2.1 Converging trends

In a broader context, one can observe two major trends in computational folkloristics [1] whose convergence will be underlying the results of the next decade. The first is focus on the evolutionary aspect of motif and/or tale type distributions, either with regard to certain tale types [44], [22], [23], [47], [4], [15] or geographical distribution of globally occurring narrative motifs [51]. Strikingly, there is a certain genetically-inspired thinking in the background, perhaps going back to the modeling capacities inherent in Dawkins' meme theory [11], comparing tale types as motif sequences to 'narrative DNA' [8], [36], [32], [34], or looking at the evolution of narrative/story networks as a quasi-biological process based on the mutation and recombination of narrative elements [23]. Such views possibly rely on certain similarities with bioinformatics in terms of network motif identification [39], a problem analog with ours. The aforementioned context is that of *evolving semantics*, an emerging research area, e.g. in digital preservation [25], [24].

The second trend is to use probabilistic and/or multivariate statistical methods for the analysis of binary or non-binary matrices of events over cases, where events can be e.g. index terms, motifs, motif sequences etc., and cases as an umbrella term stand for documents in general, e.g. abstracts describing narratives [2], tale types [52], and so on, ultimately constituting text corpora or databases. On such collections, one can then experiment with e.g. sub-corpus topic modeling (STM) by Latent Dirichlet Allocation (LDA) as a means of supervised passage exploration in partly unknown corpora [46].

The little one can say about the plethora of methods tested is that, regardless of the corpora, their regionality and the analytical units whose distributions characterize 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 [51], principal component analysis (PCA) [2], LDA [22], deep learning by recurrent neural networks (RNN) [28], or support vector machines (SVM) [35] share the same nature of being static snapshots of collections though. Of course there is an inherent contradiction in addressing text evolution, a dynamic phenomenon, by tools tailored to static measurements, but it seems to be the case that vector spaces are not really suitable to investigate semantic evolution per se, the notion asking for vector fields instead [57], [10]. Unfortunately, no semantic theory is available to explain factors behind language change or conceptual dynamics [6] in terms of vector fields for the time being.

Whereas the above approaches, and their extensions to embeddings with increasingly condensed and geometrically located types of meaning [33], [38], [42], [27], [40], [18] rely on *distributional semantics* captured by term co-occurrences, we note in passing that another method of encoding sentence semantics, reliant on *compositional semantics*, connects to quantum theory (QT) inspired text processing methods, a research direction in artificial intelligence [56]. The first publications looking at the structural study of Greek

mythology from a QT perspective were published a while ago [9], [7], expected to pave the way for similar efforts.

As the computing of results for the above both trends require datasets, we briefly look at their availability next.

2.2 Databases and datasets

Progress in computational folkloristics presupposes that experiments can be repeated by international teams to make sure that the results are robust, so that recommended methodologies can emerge based on their ranking. In order to be trusted, results must be replicable, which entails access to public datasets accessible over the Internet with as few bottlenecks as possible. Once teams of folklorists and data scientists share their respective expertise to interpret and curate the results, it will become pointless to lament about any loss of authority, or unsolvable problems in the face of scalability. It's a tea for two situation.

Moving away now from the problem of text migration combined with text evolution over millennia across the globe, below we focus on a more tractable problem. Having recently scanned the field for open access datasets of ATU-annotated tales in English as a kind of *lingua franca* (no pun intended), i.e. suitable for motif detection by machine learning, we can confirm the following:

- We could not identify such datasets on GitHub[1], Kaggle[2] or Google[3];
- Big folklore collections anticipated by [46] and [45] are still missing. Based on [31] and [21], 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 35.000 texts [46]. Other important databases exist, [3] but are either beyond public access, in their original languages only, or both. The notable exception is the Meertens Institute whose texts are of course in Dutch and Frisian plus a number of local dialects, but can be read in English translation as well. Typically, authors who work with these databases extract their research datasets but do not make them public for any number of reasons [22], [23].
- Other researchers who have shared their data as supporting material for their articles include e.g. [48], [44], [47] and [4]. Also, [13] and [12] report that recently, a large amount of ATU data have been made available online by the Multilingual Folk Tale Database (MFTD)[4], offering also annotation facilities for tales in multilingual versions. While this is a step in the right direction, the data is made available in a format which allows for browsing

[1] `https://github.com/awesomedata/awesome-public-datasets`
[2] `https://www.kaggle.com/datasets`
[3] `https://datasetsearch.research.google.com/`
[4] `https://www.mftd.org`

of portions of the database, but not for easy access to the corpus in its entirety. That being said, the English texts and related metadata contained in the MFTD might be a useful addition to the `aft` dataset.
– We found only a single recent study [28] which published a corresponding tale corpus to promote reproducibility.[5]

Among the annotated tale collections which were publicly-available on the internet, the most promising candidate was Prof D.L. Ashliman's Folktexts collection. The process of the conversion of this collection to the desired format will be described below. Importantly, Prof. Ashliman graciously agreed to donate his collection to the digital humanities (DH) and data science research communities.

3 The Ashliman Folktexts collection

The '*Folktexts*' site has been populated and maintained since 1996 by D.L. Ashliman, professor emeritus from the University of Pittsburgh. While some other sites may sport a more lavish design, Ashliman's 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. While our dataset includes only tales from pages with clear ATU annotations (n = 208 pages), the total content of the website is much larger (n = 366 pages), and includes various creation myths, stories of changelings, Faust legends, and more. It is the ATU annotation that makes this corpus particularly valuable as a 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 the Ashliman corpus, yet these often only include a smaller portion of the entire set of texts [41]. To our knowledge, none of the published studies provide an openly-accessible corpus of the data for use in promoting subsequent research.

4 Support for Reproducibility in Folklore Studies

Reproducibility is a defining characteristic of science, yet a wide gamut of scientific fields have been plagued by a "replicability crisis": a situation where trusted research findings have been impossible to reproduce [19], [37]. While the problem has come to the fore in the health and social sciences, it has been acknowledged in disciplines as broad as archaeology [29], public health [20], biology [26], and economics [30].

Reproducible research entails that study results be accompanied by:

1. a detailed description of the methods used to obtain and operate on the data;

[5] `https://github.com/GossaLo/afr-neural-folktales/`

2. the full dataset(s) used in the study;
3. the full code used to transform the data and compute the results.

In recent years the digital humanities have made some strides to emulate these efforts, with venues such as the *Journal of Open Humanities Data* being a noteworthy exception to the more common practice.

4.1 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 (CC BY-SA 4.0)[6].
- *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 (see Sect. 4.3 for additional details).
- *Open code*: Allowing any user to view and run the code that produces the dataset, as well as downstream analyses which use the dataset. This allows for inspection, refinement and reasoning about the effects of transformation and statistical modeling on the data.
- *Common form*: We have chosen to use the dataframe as the structure of the dataset, and specifically the "tidy" dataframe described by Wickham, 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 [55].
- *Common tools*: The data must also be structured in a way that allows for use with the standard tools of the trade of data science. These tools are continuously evolving, yet the dataframe is likely to continue to be common object across R (in `tidyverse`) and Python (in `pandas`). In addition, it can be read easily from a `.csv` format by Excel users to allow for ease of investigation.
- *Modifiable form*: Text analysis has traditionally used other types of data structures to model its quantitative features (e.g. document-term matrices, term co-occurrence matrices), and dataframes have been incorporated into tidy data workflows and available packages such as `quanteda` or `tidytext`. This allows for reshaping the data into sparse matrices, nested structures, and graph-based structures as dictated by the needs of a given analysis, while starting from a common source dataset (i.e. the `aft`).

[6] `https://creativecommons.org/licenses/by-sa/4.0/`

4.2 Accessing the Corpus

The current version of the `aft` dataset is located on the `trilogy` GitHub repository. For reproducible research, analyses using the dataset should use the URL identifying the current version of the dataset at the time the analysis was run.

To get a permanent link for reproducible research:

1. Navigate to the dataset's URL.
2. Press the `y` button on the keyboard to get a permanent link to the exact version of the dataset.[7] Even as the dataset changes, other researchers will be able to use this link to run their code against the same version you used.
3. Select the link to "View Raw", which will display the raw `.csv` file. The URL will change its form, and also include an access token.
4. Copy this URL. It is the key which will allow for reproducing any analysis using the dataset.

Once a permanent link has been copied, read the data into the desired environment. An example of how to access the dataset are given below using the `R` language, though similarly simple commands exist for Python and other languages:

```
aft <- read.csv("https://raw.githubusercontent.com/j-hagedorn/trilogy/.../data/aft.csv?token=...")
```

While Excel is not recommended as a tool for reproducible research, users who wish to view the file in Excel can complete steps 1-4 above and then follow Excel instructions to import data from external data sources using the same URL.

4.3 Growing and Refining the Corpus

Existing lists of motifs and tale types are incomplete [15], and thus any corpus of texts is also bound to be incomplete if its contents are limited to tales annotated using these lists.

We intend for the corpus to be supplemented over time, both through our own efforts and through the efforts of folklorists around the world. The open-source *Git* framework for collaboratively developing code provides a stable and well-documented structure for collaboratively maintaining a standard reference corpus.

A new set of annotated tales can be submitted via a "pull request" in *GitHub*, which would include a script to fetch new texts and transform them into the dataframe structure used by the `aft` dataset[8], as well as any source files required to run the script. Pull requests provide a structure for submission of changes, as well as for testing and review of newly introduced code.

[7] See the GitHub documentation on "Getting permanent links to files" for more information.

[8] Similar to the `fetch_ashliman.R` script used to obtain the initial dataset.

Collaborators are encouraged to publish descriptions of their contributions as well: for instance, as a short data paper to the *Journal of Open Humanities Data*.

In addition to growing the corpus, the *Git* framework allows for the ongoing improvement of data quality. Users of the dataset can file an issue on the repository in order to identify improvements to the data, or submit pull-requests proposing fixes to the existing scripts. This functionality allows for the corpus to serve as a communal resource, and we welcome inquiries and suggestions about how best to manage this resource as a "commons" [53].

5 Data Harvesting and Cleaning

5.1 Steps

Web-scraping of the *Folktexts* site was completed using the `rvest` package in the `R` statistical programming language. The full script is available on GitHub, and the following high-level summary of data-cleaning steps 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.[9]
2. Remove any links pointed to external websites, since these would require separate web-scraping logic to be developed.
3. Retain all links with the form `type...`, which Ashliman used to denote pages containing tales belonging to a type. Recode links which do not follow this form, but which contain tales belonging to an ATU type. For example, the page for *Animal Brides and Animal Bridegrooms* was recoded as belonging to ATU type 0402.
4. Extract the ATU type ID from the URL for each page.

The steps above result in a dataframe listing 208 webpages, each associated with a tale type and containing the page name, the page URL, and the associated ATU ID for each. This list of page URLs was looped through, using the following steps to the HTML within each page:

5. Extract HTML nodes from the page using CSS selectors (i.e. `body`, `h1`, `li` , `p`, `h3`, `a`) and create a dataframe using the text, name and attribute elements of the nodes.
6. Remove the table of contents and other superfluous text other than the tales, their titles, and other associated metadata (e.g. source documents, notes, etc.).
7. Since not all paragraphs had HTML tags, using a straightforward scraping technique would result in tales with missing sections. Therefore, we separated the `body` of each page into a separate dataframe, unnested the text

[9] The main URL for the site is `http://www.pitt.edu/~dash/folktexts.html`

by lines,[10] and used a fuzzy-joining method to align the missing body text with the well-formatted HTML.[11]
8. Join to the dataframe of extracted data elements from other URLs.

The resulting dataframe compiled the available tales from the original list of 208 webpages. To this dataframe, the following steps were applied:

9. Select the longest `text`, choosing between the tagged HTML version and the version extracted from the `body`.
10. Select the available metadata from the tagged HTML versions where those existed, using the alternate versions only if those were `NA`.
11. Remove irrelevant entries using regular expressions.
12. Create unique tale titles where these were duplicated across multiple variants of tales.
13. Clean tale text data (e.g. removing remnant HTML tags, extra spaces, replacing internal double quotes with single quotes).

5.2 Limitations

Web-scraping is an inherently messy exercise, as the data contained in web pages are often not formatted with the intent of being analyzed. Due to a broken link in the website, we were unable to obtain tales related to *The Three-Ring Parable* (0972). In addition, the pages for the following tale types were unable to be scraped, due to errors generated in the `R` session: *The Flying Dutchman, The Fool Whose Wishes All Came True, The Snow Maiden, The Strong Boy, The Tail-Fisher, What Should I Have Said (or Done)?*.

The `provenance` field does not meet the definition of "tidy" outlined above, since multiple types of descriptors (i.e. *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.

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 allow for the initial version of the dataset to be simple in its structure, and in order for machine learning to have a relatively unambiguous corpus of motif sequences to match, if using the tales as a training dataset. If required by future analyses, our intent is to store multiple ATUs as nested lists per tale within the dataframe.

6 Features of the Annotated Folktales (`aft`) dataset

[10] Using the `tidytext::unnest_tokens()` function.

[11] Using the `fuzzyjoin::stringdist_full_join()` function, we used the *Jaro-Winkler* method and set the maximum distance for a match to 1.

6.1 Data Dictionary

The `aft` (i.e. *Annotated Folktales*) dataframe contains 1559 rows, each corresponding to a single tale. Its 9 columns are described briefly below:

- `type_name` : The name associated with the Aarne-Thompson-Uther (ATU) tale type identifier.[12]
- `atu_id` : The Aarne-Thompson-Uther (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 will change as the dataset grows.
- `date_obtained` : The date on which the data set identified as a `data_source` was last downloaded and compiled.

The table below shows prints the initial characters of fields from the first 6 rows of the dataset, in order to illustrate its appearance:

atu_id	tale_title	provenance	source	text
0910b	The Highland...	Scotland	Cuthbert Bed...	In one of th...
0910b	The Prince W...	India	Cecil Henry ...	There was on...
0910b	The Three Ad...	Italy	Thomas Frede...	A man once l...
0910b	The Three Ad...	Ireland	T. Crofton C...	The stories ...
0910b	The Three Ad...	Ireland	Patrick Kenn...	The name of ...
1430	Buttermilk Jack	NA	Thomas Hughe...	Oh mother, m...

6.2 Descriptive Statistics

Length of tales. The 1559 tales in the dataset average 797 words in length, though the individual texts vary with a minimum of 11 words and a maximum of 11210. The histogram below shows the distribution of tale lengths:

[12] Note that this field currently uses the name of the associated Folktexts webpage, which may not precisely match the ATU description.

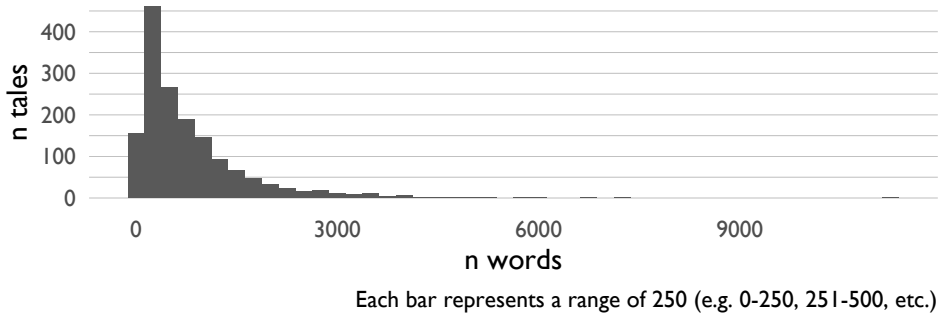

Fig. 1 Distribution of tale lengths

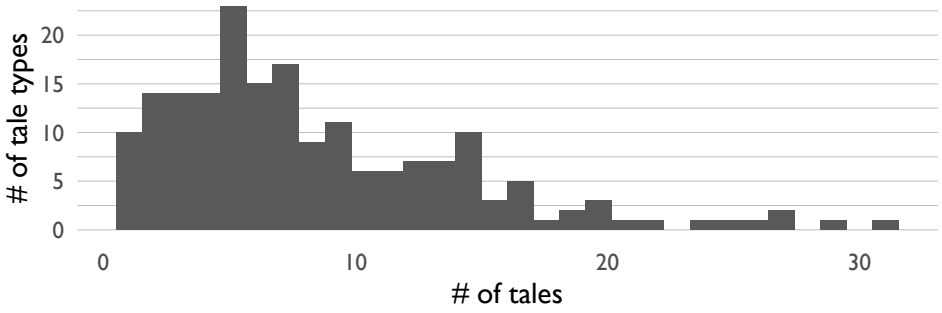

Fig. 2 Distribution of tale type membership size

Number of tales by ATU type The tales compiled in the `aft` data are annotated by Aarne-Thompson-Uther (ATU) tale type, and represent 186 distinct types. There are an average of 8.4 tales in each tale type, with a range of 1 to 31.

The tale types with the largest representative group of tales in the corpus is shown below:

7 Conclusion and Future Research

Under a Creative Commons license, we published on GitHub an open-access, ATU annotated dataset of 1559 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 assemblage to the same end. Over time it will be updated with additional information to help reproducible experimentation with supervised tale type learning.

In our upcoming work, we plan to extend the repository to provide:

– a set of all defined mythological and folktale motifs;

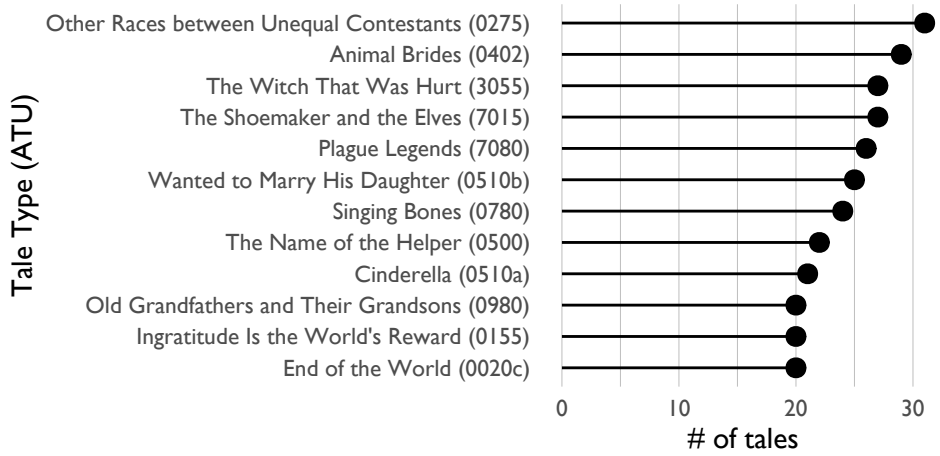

Fig. 3 Ten tale types with the largest number of representative tales

– a set of 'types', or recipes describing a sequence of motifs which are commonly used together in myths and tales;
– a collection of myth and tale texts that have been annotated as belonging to a 'type'."

References

1. Abello, J., Broadwell, P., Tangherlini, T.R.: Computational Folkloristics. Communications of the ACM **55**(7), 60–70 (2012). DOI 10.1145/2209249.2209267. URL https://doi.org/10.1145/2209249.2209267. Place: New York, NY, USA Publisher: Association for Computing Machinery
2. Berezkin, Y.: Spread of folklore motifs as a proxy for information exchange: contact zones and borderlines in Eurasia. Trames **19**(1), 3–14 (2015). URL https://pdfs.semanticscholar.org/6a90/2d19ba6d44f058c32536061d45f67a966777.pdf. Publisher: Estonian Academy Publishers
3. Berezkin, Y.E.: 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, Understanding Complex Systems, pp. 71–89. Springer International Publishing, Cham (2017). DOI 10.1007/978-3-319-39445-9_5. URL https://doi.org/10.1007/978-3-319-39445-9_5
4. 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.: Inferring patterns of folktale diffusion using genomic data. Proceedings of the National Academy of Sciences **114**(34), 9140–9145 (2017). DOI 10.1073/pnas.1614395114. URL https://www.pnas.org/content/114/34/9140
5. Darányi, S., Lendvai, P.: Proceedings of the First AMICUS Workshop, October 21, 2010 Vienna, Austria. University of Szeged, Department of Library and Human Information Science, Hungary (2010). URL http://urn.kb.se/resolve?urn=urn:nbn:se:hb:diva-3570
6. Darányi, S., Wittek, P.: Demonstrating Conceptual Dynamics in an Evolving Text Collection. Journal of the American Society for Information Science and Technology **64**(12), 2564–2572 (2013). DOI 10.1002/asi.22940. URL https://doi.org/10.1002/asi.22940

7. Darányi, S., Wittek, P.: Conceptual Machinery of the Mythopoetic Mind: Attis, A Case Study. In: H. Atmanspacher, T. Filk, E. Pothos (eds.) Quantum Interaction, Lecture Notes in Computer Science, pp. 195–206. Springer International Publishing, Cham (2016). DOI 10.1007/978-3-319-28675-4_15
8. Darányi, S., Wittek, P., Forró, L.: Toward Sequencing "Narrative DNA": Tale Types, Motif Strings and Memetic Pathways. In: Proceedings of CMN-12, 3rd Workshop on Computational Models of Narrative in conjunction with the 8th Language Resources and Evaluation Conference, pp. 2–10. Istanbul, Turkey (2012). URL `urn:nbn:se:hb:diva-6904`
9. Darányi, S., Wittek, P., Kitto, K.: The Sphynx's New Riddle: How to Relate the Canonical Formula of Myth to Quantum Interaction. In: H. Atmanspacher, E. Haven, K. Kitto, D. Raine (eds.) Quantum Interaction, Lecture Notes in Computer Science, pp. 47–58. Springer, Berlin, Heidelberg (2014). DOI 10.1007/978-3-642-54943-4_5. URL `https://link.springer.com/chapter/10.1007/978-3-642-54943-4_5`
10. Darányi, S., Wittek, P., Konstantinidis, K., Papadopoulos, S., Kontopoulos, E.: A Physical Metaphor to Study Semantic Drift. CoRR **abs/1608.01298** (2016). URL `https://arxiv.org/abs/1608.01298v1`
11. Dawkins, R.: The selfish gene. Oxford University Press (2016)
12. Declerck, T., Aman, A., Banzer, M., Macháček, D., Schäfer, L., Skachkova, N.: Multilingual ontologies for the representation and processing of folktales. pp. 20–23 (2017). DOI 10.26615/978-954-452-046-5_003
13. Declerck, T., Kostova, A., Schäfer, L.: Towards a linked data access to folktales classified by thompson's motifs and aarne-thompson-uther's types. In: Proceedings of Digital Humanities 2017. Montréal, QC, Kanada (2017). LT
14. Declerck, T., Schäfer, L.: Porting past classification schemes for narratives to a linked data framework. In: Proceedings of DATeCH2017. Göttingen (2017). LT
15. d'Huy, J., Le Quellec, J.L., Berezkin, Y., Lajoye, P., Uther, H.J.: Studying folktale diffusion needs unbiased dataset. Proceedings of the National Academy of Sciences **114**(41), E8555–E8555 (2017). DOI 10.1073/pnas.1714884114. URL `https://www.pnas.org/content/114/41/E8555`
16. Finlayson, A.M.: Inferring propp's functions from semantically annotated text. JOURNAL OF AMERICAN FOLKLORE pp. 55–77 (2016)
17. Frenzel, E.: Stoffe der Weltliteratur : ein Lexikon dichtungsgeschichtlicher Längsschnitte / Elisabeth Frenzel, 8. überarb. u. erweit. aufl. edn. Kröners Taschenausgabe ; 300. Kröner, Stuttgart (1992)
18. Garg, D., Ikbal, S., Srivastava, S.K., Vishwakarma, H., Karanam, H., Subramaniam, L.V.: Quantum embedding of knowledge for reasoning. In: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (eds.) Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc. (2019). URL `https://proceedings.neurips.cc/paper/2019/file/cb12d7f933e7d102c52231bf62b8a678-Paper.pdf`
19. Goodman, S.N., Fanelli, D., Ioannidis, J.P.A.: What does research reproducibility mean? Science Translational Medicine **8**(341), 341ps12–341ps12 (2016). DOI 10.1126/scitranslmed.aaf5027. URL `https://stm.sciencemag.org/content/8/341/341ps12`. Publisher: American Association for the Advancement of Science Section: Perspective
20. Harris, J.K., Johnson, K.J., Carothers, B.J., Combs, T.B., Luke, D.A., Wang, X.: Use of reproducible research practices in public health: A survey of public health analysts. PLoS ONE **13**(9) (2018). DOI 10.1371/journal.pone.0202447. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6135378/`
21. Ilyefalvi, E.: The Theoretical, methodological and technical issues of digital folklore databases and computational folkloristics. Acta Ethnographica Hungarica **63**(1), 209–258 (2018). URL `https://www.researchgate.net/profile/Emese-Ilyefalvi/publication/328105114_The_Theoretical_Methodological_and_Technical_Issues_of_Digital_Folklore_Databases_and_Computational_Folkloristics/links/5f045ab2a6fdcc4ca452fdc5/The-Theoretical-Methodological-and-Technical-Issues-of-Digital-Folklore-Databases-and-Computational-Folkloristics.pdf`. Publisher: Akadémiai Kiadó
22. Karsdorp, F., van den Bosch, A.: Identifying Motifs in Folktales using Topic Models, pp. 41–49 (2013). Reporting year: 2013

23. Karsdorp, F.B.: Retelling Stories: A Computational-Evolutionary Perspective. PhD Thesis, Radboud Universiteit (2016). URL http://www.karsdorp.io/retelling-stories. ISBN: 978-90-9029993-8
24. Kontopoulos, E., Darányi, S., Wittek, P., Konstantinidis, K., Riga, M., Mitzias, P., Stavropoulos, T., Andreadis, S., Maronidis, A., Karakostas, A., et al.: Deliverable 4.5: Context-aware Content Interpretation. PERICLES project (2016)
25. Kontopoulos, E., Riga, M., Mitzias, P., Andreadis, S., Stavropoulos, T., Konstantinidis, K., Maronidis, A., Karakostas, A., Tachos, S., Kaltsa, V., et al.: Pericles deliverable 4.4: modelling contextualised semantics. PERICLES project (2016)
26. Kühne, M., Liehr, A.: Improving the Traditional Information Management in Natural Sciences. Data Science Journal **8**(0), 18–26 (2009). DOI 10.2481/dsj.8.18. URL http://datascience.codata.org/articles/abstract/10.2481/dsj.8.18/. Number: 0 Publisher: Ubiquity Press
27. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. CoRR **abs/1405.4053** (2014). URL http://arxiv.org/abs/1405.4053
28. Lô, G., de Boer, V., van Aart, C.J.: Exploring West African Folk Narrative Texts Using Machine Learning. Information **11**(5) (2020). DOI 10.3390/info11050236. URL https://www.mdpi.com/2078-2489/11/5/236
29. Marwick, B.: Computational reproducibility in archaeological research: basic principles and a case study of their implementation. Faculty of Science, Medicine and Health - Papers: part A (2017). DOI 10.1007/s10816-015-9272-9. URL https://ro.uow.edu.au/smhpapers/4034
30. McCullough, B.D.: Open Access Economics Journals and the Market for Reproducible Economic Research. Economic Analysis and Policy **39**(1), 117–126 (2009). DOI 10.1016/S0313-5926(09)50047-1. URL https://www.sciencedirect.com/science/article/pii/S0313592609500471
31. Meder, T.: From a Dutch Folktale Database towards an International Folktale Database. Fabula **51**(1-2), 6–22 (2010). DOI 10.1515/FABL.2010.003. Publisher: Walter De Gruyter
32. Meder, T., Karsdorp, F., Nguyen, D.P., Theune, M., Trieschnigg, R.B., Muiser, I.: Automatic Enrichment and Classification of Folktales in the Dutch Folktale Database. The Journal of American Folklore **129**(511), 78–96 (2016). DOI 10.5406/jamerfolk.129.511.0078. Publisher: American Folklore Society
33. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013)
34. Murphy, T.P.: From Fairy Tale to Film Screenplay : Working with Plot Genotypes. Palgrave Macmillan UK, London (2015)
35. Nguyen, D., Trieschnigg, D., Meder, T., Theune, M.: Automatic classification of folk narrative genres. In: J. Jancsary (ed.) Proceedings of KONVENS 2012, pp. 378–382. ÖGAI (2012). URL http://www.oegai.at/konvens2012/proceedings/56_nguyen12w/. LThist 2012 workshop
36. Ofek, N., Darányi, S., Rokach, L.: 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, *OpenAccess Series in Informatics (OASIcs)*, vol. 32, pp. 166–182. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (2013). DOI 10.4230/OASIcs.CMN.2013.166. URL http://drops.dagstuhl.de/opus/volltexte/2013/4150. ISSN: 2190-6807
37. Pasquier, T., Lau, M.K., Trisovic, A., Boose, E.R., Couturier, B., Crosas, M., Ellison, A.M., Gibson, V., Jones, C.R., Seltzer, M.: If these data could talk. Scientific Data **4**, 170114 (2017). DOI 10.1038/sdata.2017.114
38. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (2014). DOI 10.3115/v1/D14-1162. URL https://www.aclweb.org/anthology/D14-1162
39. Qin, G., Gao, L.: An algorithm for network motif discovery in biological networks. IJDMB pp. 1–16 (2012)
40. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks (2019)

41. Reiter, N., Frank, A., Hellwig, O.: An NLP-based cross-document approach to narrative structure discovery. Literary and Linguistic Computing **29**(4), 583–605 (2014). DOI 10.1093/llc/fqu055. URL https://doi.org/10.1093/llc/fqu055
42. Rothe, S., Schütze, H.: AutoExtend: Extending word embeddings to embeddings for synsets and lexemes. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1793–1803. Association for Computational Linguistics, Beijing, China (2015). DOI 10.3115/v1/P15-1173. URL https://www.aclweb.org/anthology/P15-1173
43. Seigneuret, J.C.: Dictionary of literary themes and motifs. Greenwood Press New York (1988)
44. Silva, S.G.d., Tehrani, J.J.: Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. Royal Society open science. **3**(1), 150645 (2016). DOI 10.1098/rsos.150645. URL https://doi.org/10.1098/rsos.150645. Publisher: Royal Society
45. Tangherlini, T.: Big folklore: A special issue on computational folkloristics. The Journal of American Folklore **129**, 5 (2016). DOI 10.5406/jamerfolk.129.511.0005
46. Tangherlini, T.R., Leonard, P.: Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research. Poetics **41**(6), 725–749 (2013). DOI 10.1016/j.poetic.2013.08.002. URL https://www.sciencedirect.com/science/article/pii/S0304422X13000648
47. Tehrani, J.: The phylogeny of little red riding hood. PLoS ONE **8**(11), e78871 (2013). URL http://dro.dur.ac.uk/11481/
48. Tehrani, J.J., Nguyen, Q., Roos, T.: 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 (2016). URL http://dro.dur.ac.uk/15770/. Publisher: Oxford University Press
49. Thompson, S.: 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. edn. Rosenkilde Copenhagen (1951)
50. Thompson, S.: The Folktale. Campus (Berkeley, Calif.). University of California Press (1977). URL https://books.google.com/books?id=WKN44RtM_loC
51. Thuillard, M., d'Huy, J., Berezkin, Y., Le Quellec, J.L.: A Large-Scale Study of World Myths. Trames Journal of the Humanities and Social Sciences **22**(4), 407–424 (2018). DOI 10.3176/tr.2018.4.05. URL https://doi.org/10.3176/tr.2018.4.05
52. Uther, H.J.: The Types of International Folktales: A Classification and Bibliography, Based on the System of Antti Aarne and Stith Thompson. No. 284-286 in FF communications. Suomalainen Tiedeakatemia, Academia Scientiarum Fennica, Helsinki (2004). URL https://books.google.hu/books?id=HVQsAQAAIAAJ
53. Vollan, B., Ostrom, E.: Cooperation and the Commons. Science **330**(6006), 923–924 (2010). DOI 10.1126/science.1198349. URL https://science.sciencemag.org/content/330/6006/923. Publisher: American Association for the Advancement of Science Section: Perspective
54. White, J.: The Analysis of Music. Prentice-Hall (1976). URL https://books.google.hu/books?id=AFMkAQAAMAAJ
55. Wickham, H.: Tidy Data. Journal of Statistical Software **59**(1), 1–23 (2014). DOI 10.18637/jss.v059.i10. URL https://www.jstatsoft.org/index.php/jss/article/view/v059i10. Number: 1
56. Widdows, D., Kitto, K., Cohen, T.: Quantum mathematics in artificial intelligence. CoRR **abs/2101.04255** (2021). URL https://arxiv.org/abs/2101.04255
57. Wittek, P., Darányi, S., Kontopoulos, E., Moysiadis, T., Kompatsiaris, I.: Monitoring term drift based on semantic consistency in an evolving vector field. In: 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE, Killarney, Ireland (2015). DOI 10.1109/IJCNN.2015.7280766