Bearing a bag-of-tales: An open corpus of annotated folktales for reproducible research

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.

Table of Contents

# 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 , . There, a motif stands for a recurrent theme, whereas in musicology a motif is considered “the smallest structural unit possessing thematic identity” . In the vein of folktale research, Stith Thompson defined motifs as “the smallest element in a tale having a power to persist in tradition” .

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 . 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 , 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 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.

# 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 , or their use in ontology building . Instead, our focus will be on precursory efforts to automatic motif detection using two standard tools, the Thompson Motif Index (TMI) , and the Aarne-Thompson-Uther tale typology (ATU) . Important extensions to these, and to our current work, exist e.g. by and . Also, there exists a thin line between motifs in tales vs. motifs in myths: Walter Burkert proposed that “Myth belongs to the more general class of traditional tale” , whereas Martin P. Nilsson argued for the abundance of folktale motifs in Minoan-Mycenaean religion and its successive layers in classical Greek religion , .

## Converging trends

In a broader context, one can observe two major trends in computational folkloristics 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 , , , , , or to the geographical distribution of globally occurring narrative motifs . Strikingly, there is a certain genetically-inspired thinking in the background, perhaps going back to the modeling capacities inherent in Dawkins’ meme theory , comparing tale types as motif sequences to ‘narrative DNA’ , , , , or looking at the evolution of narrative/story networks as a quasi-biological process based on the mutation and recombination of narrative elements . Such views possibly rely on certain similarities with bioinformatics in terms of network motif identification , a problem analog with ours. The aforementioned context is that of *evolving semantics*, an emerging research area, e.g. in digital preservation , .

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 , tale types , 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 .

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 , principal component analysis (PCA) , LDA , deep learning by recurrent neural networks (RNN) , or support vector machines (SVM) 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 , . Unfortunately, no semantic theory is available to explain factors behind language change or conceptual dynamics 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 , , , , , 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 . The first publications looking at the structural study of Greek mythology from a QT perspective were published a while ago , , 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.

## 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]](#footnote-22), Kaggle[[2]](#footnote-24) or Google[[3]](#footnote-26);
* Big folklore collections anticipated by and are still missing. Based on and , 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 . Other important databases exist, but are either beyond public access, or 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 , .
* Other researchers who have shared their data as supporting material for their articles include e.g. , , and . Also, and report that recently, a large amount of ATU data have been made available online by the Multilingual Folk Tale Database (MFTD)[[4]](#footnote-28), offering also annotation facilities for tales in multilingual versions. While this is a step in the right direction, the data are made available in a format which allows for browsing of portions of the database, but not for easy access to the corpus in its entirety.
* We found only a single recent study which published a corresponding tale corpus to promote reproducibility.[[5]](#footnote-30)

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 and data science research communities.

# 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 = 212 pages), the total content of the website is much larger (n = 368 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 . To our knowledge, none of the published studies provide an openly-accessible corpus of the data for use in promoting subsequent research.

# 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 , . While the problem has come to the fore in the health and social sciences, it has been acknowledged in disciplines as broad as archaeology , public health , biology , and economics .

Reproducible research entails that study results be accompanied by:

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.

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.

## 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]](#footnote-35).
* *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*: 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 .
* *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).

## Accessing and Growing the Corpus

The current version of the aft dataset is located on the trilogy GitHub repository, where a vignette[[7]](#footnote-38) provides information on how to access the dataset, use it in reproducible research, and improve it.

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

# Data Harvesting and Cleaning

## 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.[[8]](#footnote-42)
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 212 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:

1. 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.
2. 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.).
3. 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 by lines,[[9]](#footnote-43) and used a fuzzy-joining method to align the missing body text with the well-formatted HTML.[[10]](#footnote-44)
4. Join to the dataframe of extracted data elements from other URLs.

The resulting dataframe compiled the available tales from the original list of 212 webpages. To this dataframe, the following steps were applied:

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

## 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* (0920e). In addition, the pages for the following tale types were unable to be scraped, due to errors generated in the R session: *The Fisherman and His Wife, A Fool Does Not Count the Animal He Is Riding, Singing Thieves, Stages of Life, Straw, Bean, and Coal, War between the Village Animals and the Forest Animals*.

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.

# Features of the Annotated Folktales (aft) dataset

## 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:

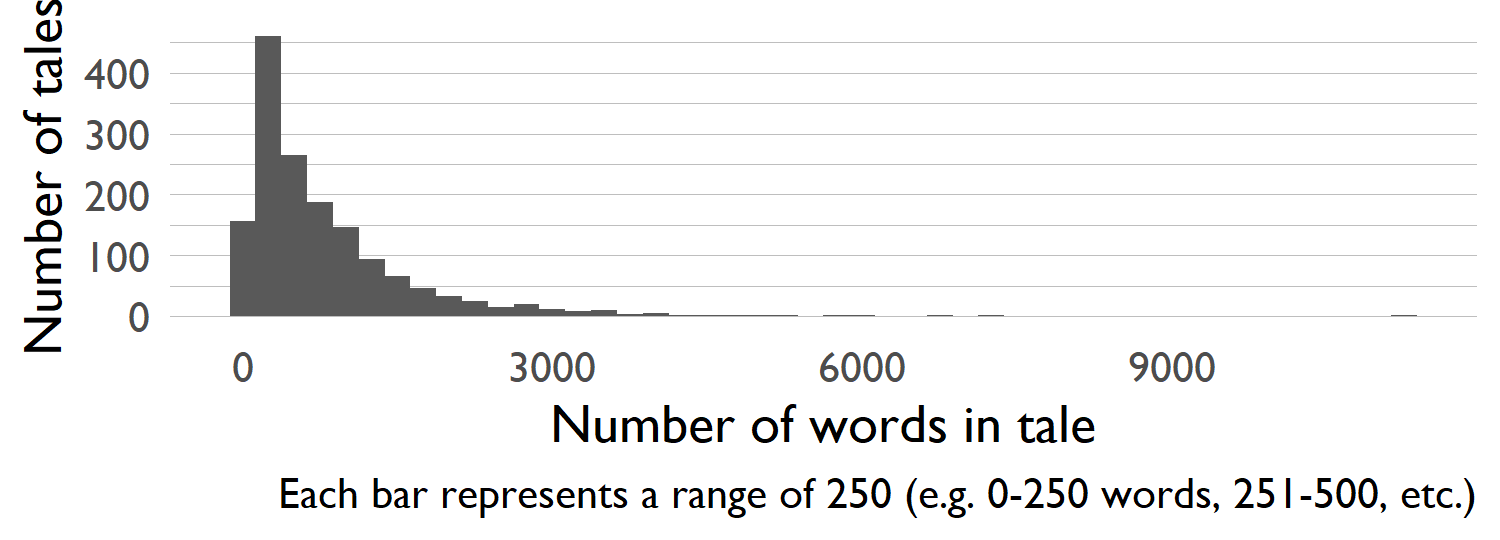
* type\_name : The name associated with the Aarne-Thompson-Uther (ATU) tale type identifier.[[11]](#footnote-48)
* atu\_id : The Aarne-Thompson-Uther (ATU) tale type identifier which classifies the tale.
* tale\_title : The title of the tale.
* provenance : The person, place or tradition from which the 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:

## Descriptive Statistics

#### Length of tales.

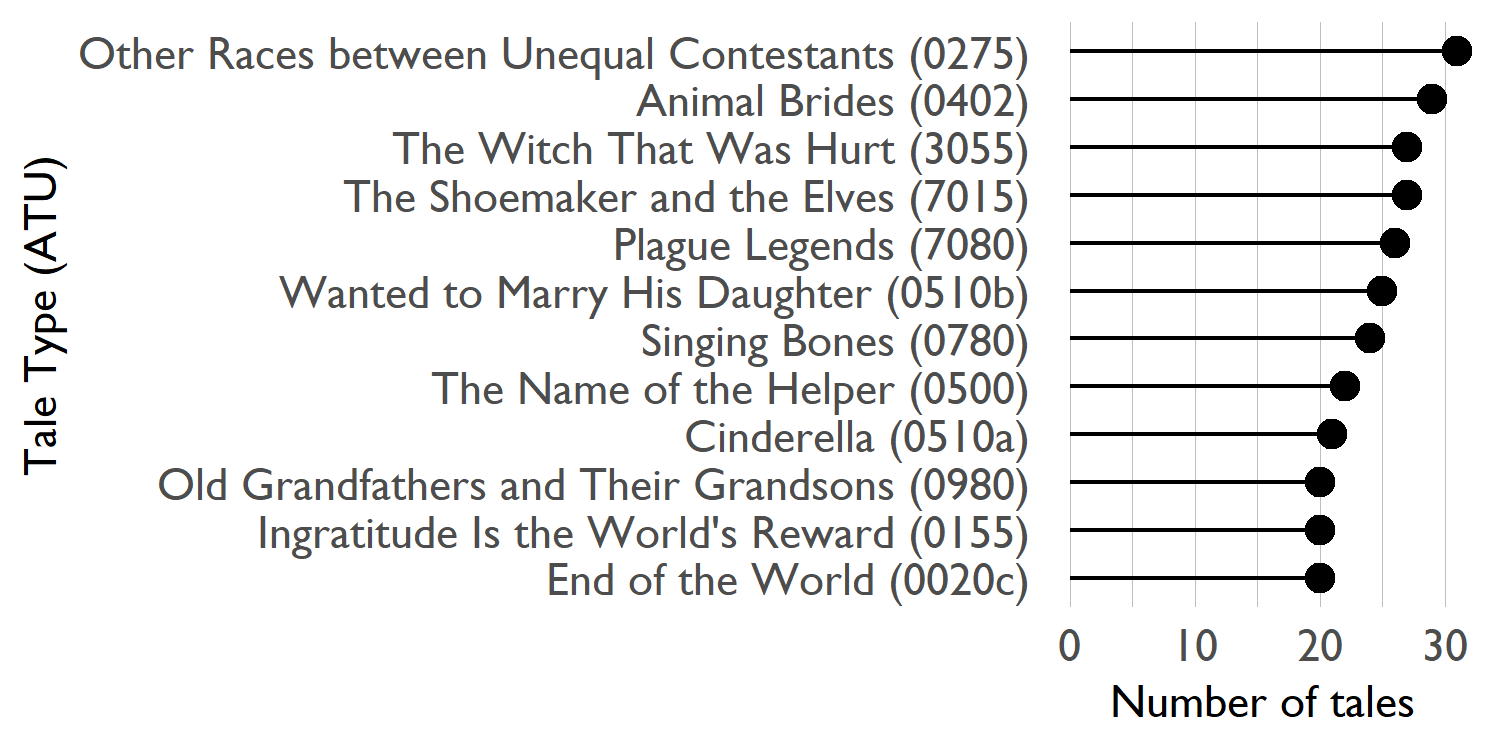
The 1559 tales in the dataset average 797 words in length, though the individual texts vary with a minimum of 11 words and a maximum of 11210. The histogram below shows the distribution of tale lengths:



Distribution of tale lengths

#### 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:



Ten tale types with the largest number of representative tales

# 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;
* 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’.”

# Acknowledgements

We thank Professor D.L. Ashliman (University of Pittsburgh) for his permission to use the data from his annotated Folktexts as a publicly available resource.

# References

1. <https://github.com/awesomedata/awesome-public-datasets> [↑](#footnote-ref-22)
2. <https://www.kaggle.com/datasets> [↑](#footnote-ref-24)
3. <https://datasetsearch.research.google.com/> [↑](#footnote-ref-26)
4. <https://www.mftd.org> [↑](#footnote-ref-28)
5. <https://github.com/GossaLo/afr-neural-folktales/> [↑](#footnote-ref-30)
6. <https://creativecommons.org/licenses/by-sa/4.0/> [↑](#footnote-ref-35)
7. <https://github.com/j-hagedorn/trilogy/blob/master/docs/vignettes/getting_started.md> [↑](#footnote-ref-38)
8. The main URL for the site is http://www.pitt.edu/~dash/folktexts.html [↑](#footnote-ref-42)
9. Using the tidytext::unnest\_tokens() function. [↑](#footnote-ref-43)
10. Using the fuzzyjoin::stringdist\_full\_join() function, we used the *Jaro-Winkler* method and set the maximum distance for a match to 1. [↑](#footnote-ref-44)
11. Note that this field currently uses the name of the associated Folktexts webpage, which may not precisely match the ATU description. [↑](#footnote-ref-48)