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

A bag-of-tales from Santa

Converting the Ashliman Folktexts Collection into a dataset for machine learning

Sándor Darányi · Joshua Hagedorn ·

Received: date / Accepted: date

Abstract Computational motif identification in folktales is an open research problem. To move ahead in this area, the field would benefit from shared test data for machine learning, putting experimentation in focus. Folklore databases including text collections in multiple languages do exist, but not in dataset form for data science, and are currently not shared, making their results non-reproducible, an obstacle to scientific progress. The need for significant preprocessing adds insult to injury, rendering the outcome both incomparable and subject to multidisciplinary criticism. As a first step to remedy this problem, we report work in progress, having converted the Ashliman Folktexts Collection into a public dataset for supervised tale type learning, itself a precondition for scalable motif identification. In the future, this dataset can be upgraded in several respects to serve as the basis for springboard experiments with the Thompson Motif Index and the Aarne-Thompson-Uther tale typology, paving the way for ontology development.

Keywords folktale · mythology · motif · reproducibility · machine learning ·

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.

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

Joshua Hagedorn
Department of ZZZ, University of WWW
E-mail: josh.hagedorn@gmail.com

1 Introduction

Ever since the concept of a motif was introduced some 200 years ago, the quest to identify content elements above 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 motive is considered "the smallest structural unit possessing thematic identity" [White(1976)]. In a similar vein, Stith Thompson defined motifs in folktale research as "the smallest element in a tale having a power to persist in tradition" [Thompson(1977)] (1946).

A sufficient overlap between these definitions suggests that such higher order content units exist as narrative building blocks in a generic sense, but their automatic extraction by computational means has eluded folk narrative studies so far [Darányi and Lendvai(2010)]. In spite of 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 [Karsdorp and van den Bosch(2013)], 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 [Thompson(1951)] 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 take a small 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 [Finlayson(2016)], or their use in ontology building [Declerck et al(2017a)Declerck, Aman, Banzer, Macháček, Schäfer, and Skachkova]. Instead, our focus will be on precursory efforts to automatic motif detection using two standard tools, the Thompson Motif Index (TMI) [Thompson(1951)], and the Aarne-Thompson-Uther tale typology (ATU) [Uther and Fellows(2004)]. Important extensions to these, and to our current work, exist e.g. by [Declerck and Schäfer(2017)] and [Declerck et al(2017b)Declerck, Kostova, and Schäfer] .

2.1 Converging trends

In a broader context, one can observe two major trends in computational folkloristics [Abello et al(2012)Abello, Broadwell, and Tangherlini] 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 [Silva and Tehrani(2016)], [Karsdorp and van den Bosch(2013)], [Karsdorp(2016)], [Tehrani(2013)], [Bortolini et al(2017)Bortolini, Pagani, Crema, Sarno, Barbieri, Boattini, Sazzini, da S [d'Huy et al(2017)d'Huy, Le Quellec, Berezkin, Lajoye, and Uther] or geographical distribution of globally occurring narrative motifs [Thuillard et al(2018)Thuillard, d'Huy, Berezkin, and Le Quellec]. Strikingly, there is a certain genetically-inspired thinking in the background, perhaps going back to the modeling capacities inherent in Dawkins' meme theory [Dawkins(2016)], comparing tale types as motif sequences to 'narrative DNA' [Darányi et al(2012)Darányi, Wittek, and Forró], [Ofek et al(2013)Ofek, Darányi, and Rokach], [Meder et al(2016)Meder, Karsdorp, Nguyen, Theune, Trieschnigg, and Muiser], [Murphy(2015)], or looking at the evolution of narrative/story networks as a quasi-biological process based on the mutation and recombination of narrative elements [Karsdorp(2016)]. Such views possibly rely on certain similarities with bioinformatics in terms of network motif identification [Qin and Gao(2012)], a problem analog with ours. The aforementioned context is that of *evolving semantics*, an emerging research area, e.g. in digital preservation [Kontopoulos et al(2016b)Kontopoulos, Riga, Mitzias, And [Kontopoulos et al(2016a)Kontopoulos, Darányi, Wittek, Konstantinidis, Riga, Mitzias, Stavropoulos, Andreadis, Maron

The second trend is to use probabilistic and/or multivariate statistical methods for the analysis of binary or non-binary co-occurrence 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 (Berezkin 2015b), tale types [Uther and Fellows(2004)], 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 (Tangherlini and Leonard 2013 Trawling).

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 [Thuillard et al(2018)Thuillard, d'Huy, Berezkin, and Le Quellec], principal component analysis (PCA) (Berezkin 2015b), LDA [Karsdorp and van den Bosch(2013)], deep learning by recurrent neural networks (RNN) [Lô et al(2020)Lô, de Boer, and van Aart], support vector machines (SVM) [Nguyen et al(2012)Nguyen, Trieschnigg, Meder, and Theune] 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 [Wittek et al(2015)Wittek, Darányi, Kontopoulos, Moysiadis, and Kompatsia [Darányi et al(2016)Darányi, Wittek, Konstantinidis, Papadopoulos, and Kontopoulos]. Unfortunately, no semantic theory is available to explain factors behind language change or conceptual dynamics [Darányi and Wittek(2013)] 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 [Mikolov et al(2013)Mikolov, Chen, Corrado, and Dean], [Pennington et al(2014)Pennington, Socher, and Manning], [Rothe and Schütze(2015)], [Le and Mikolov(2014)], [Reimers and Gurevych(2019)], [Garg et al(2019)Garg, Ikbal, Srivastava, Vishwakarma, Karanar 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 [Widdows et al(2021)Widdows, Kitto, and Cohen]. The first publications looking at the structural study of Greek mythology from a QT perspective were published a while ago [Darányi et al(2014)Darányi, Wittek, and Kitto], [Darányi and Wittek(2016)], 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

D'Huy et al with U. Datasets extracted from databases must exist but are not published. Berezhkin dataset in Russian only. This is a catch-22 situation: A Dutch will never repeat the experiment and a non-Dutch will never be able to do so. The same holds for Russian, Estonian, Hungarian, etc. The closest to a lingua franca, no pun intended, is to default on English. GS survey returns practically nil. Meder survey, Ilyefalvi, all articles rely on ones of own manufacturing, plus neither are in the public domain. Evolving datasets even less so (Karsdorp 2016). One of the exceptions that qualified in every respect, and was graciously donated to the digital humanities (DH) and data science community, is Prof DL Ashliman's Folktexts (http).

Tangherlini 2016: Big folklore implies pattern discovery at large but the respective datasets are nowhere to be accessed.

Berezkin 2015b states that only the catalog is placed on the web but not the corresponding files. Research potential thereafter is nominal at best.Site was promised to be opened but I wonder. Cca 50000 abstracts, comparable with Meertens, but short of institutional support, in Russian only.

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 we have only included tales from pages with clear ATU annotations (n = 208 pages) in this dataset, the total content of the website is much larger (n = 366 pages), and includes various creation myths, stories of changelings, Faust legends, and more.

Despite the richness of this resource, it has not frequently been used in folklore research as a larger corpus. While some previous studies reference the Ashliman corpus, these often only include a smaller portion of the entire set of texts [Reiter et al(2014)Reiter, Frank, and Hellwig]. 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 [Goodman et al(2016)Goodman, Fanelli, and Ioannidis], [Pasquier et al(2017)Pasquier, Lau, Trisovic, Boose, Couturier, Crosas, Ellison, Gibson, Jones, and Seltzer]. 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)Harris, Johnson, Carothers, Combs, Luke, and Wang], biology [Kühne and Liehr(2009)], and economics [McCullough(2009)].

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.

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 license [cite].
- *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. 4.2 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 [Wickham(2014)].
- *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`).

4.2 Growing the Corpus

- motifs, tale types and tale corpus are incomplete, but that does not mean they should be thrown out
- need structure for adding new tales
- pull request provides structure for submission and review of changes
- this can also be used to identify and correct errors (so publish and PR)
- for reproducible research, articles using the datasets should use the url with the current commit's SHA to indicate the state of the dataset at the time the analysis was run.

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.[1]
2. Remove any links pointed to external websites, since these would require separate web-scraping logic to be developed.

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

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 by lines,[2] and used a fuzzy-joining method to align the missing body text with the well-formatted HTML.[3]
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

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

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

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

6.1 Data Dictionary

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

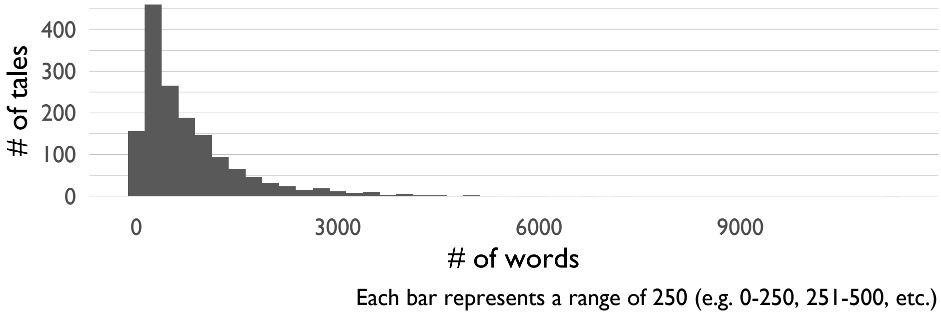
- type_name : The name associated with the Aarne-Thompson-Uther (ATU) tale type identifier.
- atu_id : The Aarne-Thompson-Uther (ATU) tale type identifier which classifies the tale.
- tale_title : The title of the tale.
- provenance : The person, place or tradition from which the tale came. In Ashliman's 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:

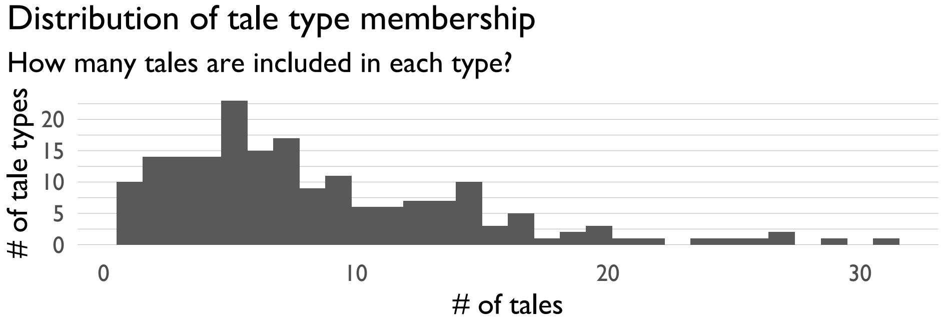
atu_id	tale_title	provenance	source	text
0910b	The Highland...	Scotland	Cuthbert Bed...	In one of th...
0910b	The Prince W...	India	Cecil Henry ...	There was on...
0910b	The Three Ad...	Italy	Thomas Frede...	A man once l...
0910b	The Three Ad...	Ireland	T. Crofton C...	The stories ...
0910b	The Three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

6.2 Descriptive Statistics

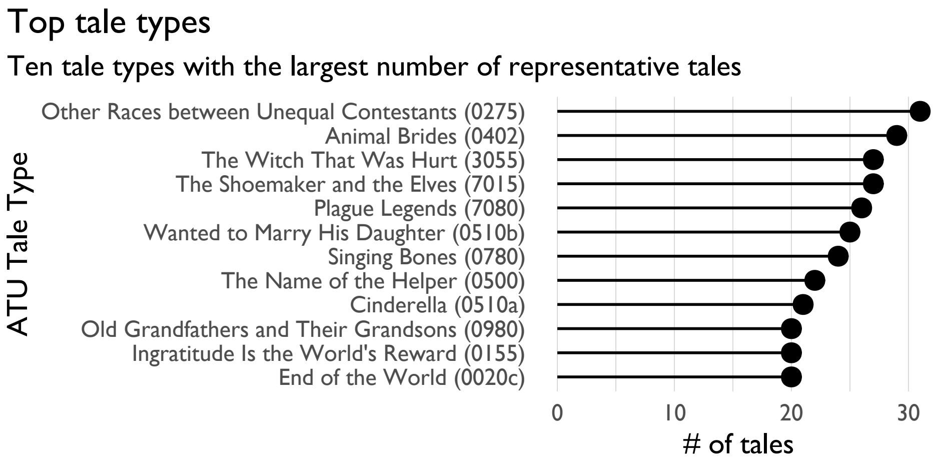
Length of tales. The 1559 tales in the dataset average 797 words in length, though the individual texts vary with a minimum of 11 words and a maximum of 11210. The histogram below shows the distribution of tale lengths:



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

7.1 Future

The intent of the repository is 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'

References

Abello et al(2012)Abello, Broadwell, and Tangherlini. Abello J, Broadwell P, Tangherlini TR (2012) Computational Folkloristics. Communications of the ACM 55(7):60–70, DOI 10.1145/2209249.2209267, URL https://doi.org/10.1145/2209249.2209267, place: New York, NY, USA Publisher: Association for Computing Machinery

Bortolini et al(2017)Bortolini, Pagani, Crema, Sarno, Barbieri, Boattini, Sazzini, da Silva, Martini, Metspalu, Pettener, Luiselli, and Tehrani. Bortolini E, Pagani L, Crema ER, Sarno S, Barbieri C, Boattini A, Sazzini M, da Silva SG, Martini G, Metspalu M, Pettener D, Luiselli D, Tehrani JJ (2017) Inferring patterns of folktale diffusion using genomic data. Proceedings of the National Academy of Sciences 114(34):9140–9145, DOI 10.1073/pnas.1614395114, URL https://www.pnas.org/content/114/34/9140, https://www.pnas.org/content/114/34/9140.full.pdf

Darányi and Lendvai(2010). Darányi S, Lendvai P (2010) Proceedings of the First AMICUS Workshop, October 21, 2010 Vienna, Austria. University of Szeged, Department of Library and Human Information Science, Hungary, URL http://urn.kb.se/resolve?urn=urn:nbn:se:hb:diva-3570

Darányi and Wittek(2013). Darányi S, Wittek P (2013) Demonstrating Conceptual Dynamics in an Evolving Text Collection. Journal of the American Society for Information Science and Technology 64(12):2564–2572, DOI 10.1002/asi.22940, URL https://doi.org/10.1002/asi.22940

Darányi and Wittek(2016). Darányi S, Wittek P (2016) Conceptual Machinery of the Mythopoetic Mind: Attis, A Case Study. In: Atmanspacher H, Filk T, Pothos E (eds) Quantum Interaction, Springer International Publishing, Cham, Lecture Notes in Computer Science, pp 195–206, DOI 10.1007/978-3-319-28675-4_15

Darányi et al(2012)Darányi, Wittek, and Forró. Darányi S, Wittek P, Forró L (2012) 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, Istanbul, Turkey, pp 2–10, URL urn:nbn:se:hb:diva-6904

Darányi et al(2014)Darányi, Wittek, and Kitto. Darányi S, Wittek P, Kitto K (2014) The Sphynx's New Riddle: How to Relate the Canonical Formula of Myth to Quantum Interaction. In: Atmanspacher H, Haven E, Kitto K, Raine D (eds) Quantum Interaction, Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, pp 47–58, DOI 10.1007/978-3-642-54943-4_5, URL https://link.springer.com/chapter/10.1007/978-3-642-54943-4_5

Darányi et al(2016)Darányi, Wittek, Konstantinidis, Papadopoulos, and Kontopoulos. Darányi S, Wittek P, Konstantinidis K, Papadopoulos S, Kontopoulos E (2016) A Physical Metaphor to Study Semantic Drift. CoRR abs/1608.01298, URL https://arxiv.org/abs/1608.01298v1

Dawkins(2016). Dawkins R (2016) The selfish gene. Oxford University Press

Declerck and Schäfer(2017). Declerck T, Schäfer L (2017) Porting past classification schemes for narratives to a linked data framework. In: Proceedings of DATeCH2017, Göttingen, lT

Declerck et al(2017a)Declerck, Aman, Banzer, Macháček, Schäfer, and Skachkova. Declerck T, Aman A, Banzer M, Macháček D, Schäfer L, Skachkova N (2017a) Multilingual ontologies for the representation and processing of folktales. pp 20–23, DOI 10.26615/978-954-452-046-5_003

Declerck et al(2017b)Declerck, Kostova, and Schäfer. Declerck T, Kostova A, Schäfer L (2017b) 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, lT

d'Huy et al(2017)d'Huy, Le Quellec, Berezkin, Lajoye, and Uther. d'Huy J, Le Quellec JL, Berezkin Y, Lajoye P, Uther HJ (2017) Studying folktale diffusion needs unbiased dataset. Proceedings of the National Academy of Sciences 114(41):E8555–E8555, DOI 10.1073/pnas.1714884114, URL https://www.pnas.org/content/114/41/E8555, https://www.pnas.org/content/114/41/E8555.full.pdf

Finlayson(2016). Finlayson AM (2016) Inferring propp's functions from semantically annotated text. JOURNAL OF AMERICAN FOLKLORE pp 55–77

Frenzel(1992). Frenzel E (1992) Stoffe der Weltliteratur : ein Lexikon dichtungsgeschichtlicher Längsschnitte / Elisabeth Frenzel, 8th edn. Kröners Taschenausgabe ; 300, Kröner, Stuttgart

Garg et al(2019)Garg, Ikbal, Srivastava, Vishwakarma, Karanam, and Subramaniam. Garg D, Ikbal S, Srivastava SK, Vishwakarma H, Karanam H, Subramaniam LV (2019) Quantum embedding of knowledge for reasoning. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, URL https://proceedings.neurips.cc/paper/2019/file/cb12d7f933e7d102c52231bf62b8a678-Paper.pdf

Goodman et al(2016)Goodman, Fanelli, and Ioannidis. Goodman SN, Fanelli D, Ioannidis JPA (2016) What does research reproducibility mean? Science Translational Medicine 8(341):341ps12–341ps12, DOI 10.1126/scitranslmed.aaf5027, URL https://stm.sciencemag.org/content/8/341/341ps12, publisher: American Association for the Advancement of Science Section: Perspective

Harris et al(2018)Harris, Johnson, Carothers, Combs, Luke, and Wang. Harris JK, Johnson KJ, Carothers BJ, Combs TB, Luke DA, Wang X (2018) Use of reproducible research practices in public health: A survey of public health analysts. PLoS ONE 13(9), DOI 10.1371/journal.pone.0202447, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6135378/

Karsdorp and van den Bosch(2013). Karsdorp F, van den Bosch A (2013) Identifying Motifs in Folktales using Topic Models, pp 41–49. Reporting year: 2013

Karsdorp(2016). Karsdorp FB (2016) Retelling Stories: A Computational-Evolutionary Perspective. PhD Thesis, Radboud Universiteit, URL http://www.karsdorp.io/retelling-stories, iSBN: 978-90-9029993-8

Kontopoulos et al(2016a)Kontopoulos, Darányi, Wittek, Konstantinidis, Riga, Mitzias, Stavropoulos, Andreadis, Maronidis, Karakostas et al. Kontopoulos E, Darányi S, Wittek P, Konstantinidis K, Riga M, Mitzias P, Stavropoulos T, Andreadis S, Maronidis A, Karakostas A, et al (2016a) Deliverable 4.5: Context-aware Content Interpretation. PERICLES project

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

Kühne and Liehr(2009). Kühne M, Liehr A (2009) Improving the Traditional Information Management in Natural Sciences. Data Science Journal 8(0):18–26, DOI 10.2481/dsj.8.18, URL http://datascience.codata.org/articles/abstract/10.2481/dsj.8.18/, number: 0 Publisher: Ubiquity Press

Le and Mikolov(2014). Le QV, Mikolov T (2014) Distributed representations of sentences and documents. CoRR abs/1405.4053, URL http://arxiv.org/abs/1405.4053, 1405.4053

Lô et al(2020)Lô, de Boer, and van Aart. Lô G, de Boer V, van Aart CJ (2020) Exploring West African Folk Narrative Texts Using Machine Learning. Information 11(5), DOI 10.3390/info11050236, URL https://www.mdpi.com/2078-2489/11/5/236

Marwick(2017). Marwick B (2017) Computational reproducibility in archaeological research: basic principles and a case study of their implementation. Faculty of Science, Medicine and Health - Papers: part A DOI 10.1007/s10816-015-9272-9, URL https://ro.uow.edu.au/smhpapers/4034

McCullough(2009). McCullough BD (2009) Open Access Economics Journals and the Market for Reproducible Economic Research. Economic Analysis and Policy 39(1):117–126, DOI 10.1016/S0313-5926(09)50047-1, URL https://www.sciencedirect.com/science/article/pii/S0313592609500471

Meder et al(2016)Meder, Karsdorp, Nguyen, Theune, Trieschnigg, and Muiser. Meder T, Karsdorp F, Nguyen DP, Theune M, Trieschnigg RB, Muiser I (2016) Automatic Enrichment and Classification of Folktales in the Dutch Folktale Database. The Journal of American Folklore 129(511):78–96, DOI 10.5406/jamerfolk.129.511.0078, publisher: American Folklore Society

Mikolov et al(2013)Mikolov, Chen, Corrado, and Dean. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. 1301.3781

Murphy(2015). Murphy TP (2015) From Fairy Tale to Film Screenplay : Working with Plot Genotypes. Palgrave Macmillan UK, London

Nguyen et al(2012)Nguyen, Trieschnigg, Meder, and Theune. Nguyen D, Trieschnigg D, Meder T, Theune M (2012) Automatic classification of folk narrative genres. In: Jancsary J (ed) Proceedings of KONVENS 2012, ÖGAI, pp 378–382, URL http://www.oegai.at/konvens2012/proceedings/56_nguyen12w/, lThist 2012 workshop

Ofek et al(2013)Ofek, Darányi, and Rokach. Ofek N, Darányi S, Rokach L (2013) Linking Motif Sequences with Tale Types by Machine Learning. In: Finlayson MA, Fisseni B, Löwe B, Meister JC (eds) 2013 Workshop on Computational Models of Narrative, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, OpenAccess Series in Informatics (OASIcs), vol 32, pp 166–182, DOI 10.4230/OASIcs.CMN.2013.166, URL http://drops.dagstuhl.de/opus/volltexte/2013/4150, iSSN: 2190-6807

Pasquier et al(2017)Pasquier, Lau, Trisovic, Boose, Couturier, Crosas, Ellison, Gibson, Jones, and Seltzer. Pasquier T, Lau MK, Trisovic A, Boose ER, Couturier B, Crosas M, Ellison AM, Gibson V, Jones CR, Seltzer M (2017) If these data could talk. Scientific Data 4:170114, DOI 10.1038/sdata.2017.114

Pennington et al(2014)Pennington, Socher, and Manning. Pennington J, Socher R, Manning C (2014) GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, pp 1532–1543, DOI 10.3115/v1/D14-1162, URL https://www.aclweb.org/anthology/D14-1162

Qin and Gao(2012). Qin G, Gao L (2012) An algorithm for network motif discovery in biological networks. IJDMB pp 1–16

Reimers and Gurevych(2019). Reimers N, Gurevych I (2019) Sentence-bert: Sentence embeddings using siamese bert-networks. 1908.10084

Reiter et al(2014)Reiter, Frank, and Hellwig. 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 10.1093/llc/fqu055, URL https://doi.org/10.1093/llc/fqu055

Rothe and Schütze(2015). Rothe S, Schütze H (2015) 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), Association for Computational Linguistics, Beijing, China, pp 1793–1803, DOI 10.3115/v1/P15-1173, URL https://www.aclweb.org/anthology/P15-1173

Seigneuret(1988). Seigneuret JC (1988) Dictionary of literary themes and motifs. Greenwood Press New York

Silva and Tehrani(2016). Silva SGd, Tehrani JJ (2016) Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. Royal Society open science 3(1):150645, DOI 10.1098/rsos.150645, URL https://doi.org/10.1098/rsos.150645, publisher: Royal Society

Tehrani(2013). Tehrani J (2013) The phylogeny of little red riding hood. PLoS ONE 8(11):e78871, URL http://dro.dur.ac.uk/11481/

Thompson(1951). 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. edn. Rosenkilde Copenhagen

Thompson(1977). Thompson S (1977) The Folktale. Campus (Berkeley, Calif.), University of California Press, URL https://books.google.com/books?id=WKN44RtM_loC

Thuillard et al(2018)Thuillard, d'Huy, Berezkin, and Le Quellec. Thuillard M, d'Huy J, Berezkin Y, Le Quellec JL (2018) A Large-Scale Study of World Myths. Trames Journal of the Humanities and Social Sciences 22(4):407–424, DOI 10.3176/tr.2018.4.05, URL https://doi.org/10.3176/tr.2018.4.05

Uther and Fellows(2004). Uther HJ, Fellows F (2004) 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, URL https://books.google.hu/books?id=HVQsAQAAIAAJ

White(1976). White J (1976) The Analysis of Music. Prentice-Hall, URL https://books.google.hu/books?id=AFMkAQAAMAAJ

Wickham(2014). Wickham H (2014) Tidy Data. Journal of Statistical Software 59(1):1–23, DOI 10.18637/jss.v059.i10, URL https://www.jstatsoft.org/index.php/jss/article/view/v059i10, number: 1

Widdows et al(2021)Widdows, Kitto, and Cohen. Widdows D, Kitto K, Cohen T (2021) Quantum mathematics in artificial intelligence. CoRR abs/2101.04255, URL https://arxiv.org/abs/2101.04255, 2101.04255

Wittek et al(2015)Wittek, Darányi, Kontopoulos, Moysiadis, and Kompatsiaris. Wittek P, Darányi S, Kontopoulos E, Moysiadis T, Kompatsiaris I (2015) Monitoring term drift based on semantic consistency in an evolving vector field. In: 2015 International Joint Conference on Neural Networks (IJCNN), IEEE, Killarney, Ireland, pp 1–8, DOI 10.1109/IJCNN.2015.7280766