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: abc@def

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 1976, 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 the power to persist in tradition" (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 (Daránvi 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 (TMI, 1955-58) 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 scalable pattern hunt yet.

Our research problem for the current paper is this: consider the case of two standard reference tools, the TMI, and the Aarne-Thompson-Uther tale type index (ATU, 2004). A count in the TMI indicated the presence of .... motifs, whereas Yarlott and Finlayson counted 46,248 motifs and sub-motifs from over 614 collections, 41,796 of which had references to tales or tale types (2016). However, based on our count the ATU uses only .... (... %) of them to model tale structures as motif strings. One is then prompted to ask, where have .... % of motifs in the TMI disappeared, and how can an important monograph acquire almost canonical status with such a discrepancy in its background? In our eyes, the explanation may go back to the very different comparison capacities of the human mind vs. the computer, leading to differently robust deductions, and for a remedy to this situation one needs to call in data science. Namely 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 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

Instead of enumerating important steps in computational folkloristics (Abello et al 2012) in a chronological manner, for the sake of brevity we present related progress in the context of evolving semantics as the subject in study. In interaction with vector space based methods to represent increasingly condensed semantic content by means of embeddings, and with predication based semantic indexing (Cohen and Widdows), this emerging new paradigm can be observed e.g. in digital preservation (D4.4, D4.5). In computational folk narrative studies, two trends are converging in this broader context.

2.1 Converging trends

Those interested in the nature of progress in the field of computational folkloristics can observe two major trends whose convergence will be underlying the results of the next decade. The first of these is focus on the evolutionary aspect of motif and/or tale type distributions, either with regard to certain tale types (eg Tehrani, d'Huy and Tehrani, Karsdorp and van den Bosch 2013, Karsdorp 2016), or geographical distribution of globally occurring motifs (Berezhkin, d'Huy), or both (refs). Strikingly, there is a certain genetically inspired thinking in the background ultimately going back to the modelling capacities inherent in Dawkins' meme theory (ref), comparing tale types as motif sequences to 'narrative DNA' (Darányi et al 2012, Ofek et al, Meder MUSE), or looking at the evolution of narrative/storytelling networks as a quasi-biological process based on mutation (Karsdorp). Such views must have been informed by certain similarities with bioinformatics in terms of network motif identification (ref), a problem analog with ours. An early example of this idea was based on the recombination of narrative elements as its driving mechanism, unfortunately available in Hungarian only (Darányi), leading to another publication (Voigt et al 1999).

The second major trend one can observe, that of computing embeddings [define embedding] of increasingly condensed word and sentence semantics and beyond (refs), must be familiar from information retrieval, machine learning, data science or knowledge representation. Many computational folkloristics results manifest early examples of this paradigm (Darányi Infonautica, Tangherlini, West African, etc). One aspect of such methods is that they share distributional semantics for a default, and locate content in vector space.

Significantly, another method of sentence semantics encoding connects to quantum theory (QT) inspired text processing methods, another research direction in artificial intelligence (Widdows). On a parallel track, the first publications looking at mythologies from a QT perspective were published a while ago (CF, Attis), expected to cross-pollinate current practice. However, vector spaces are not really suitable to investigate semantic evolution in any respect, the concept asking for vector fields to model the inherent dynamics of content (arxiv 1, 2). Unfortunately no semantic theory is available to explain factors

behind language change or conceptual dynamics (JASIST) in terms of vector fields for the time being.

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).

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 have 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 = 126 pages), the total content of the website is much larger (n = 216 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.

While some previous studies reference the corpus, these often only include a smaller portion of the entire set of texts [Reiter et al(2014)Reiter, Frank, and Hellwig].

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 [cite]. While the problem has come to the fore in the health and social sciences, it has been

acknowledged in disciplines as broad as archaeology [cite], political science [cite], biology [cite], and economics [cite].

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 some strides have been made in the digital humanities to emulate these efforts, with 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 AFT 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.
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 126 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

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

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 126 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

- Unable to scrape broken links
- Following pages from the initial set of URLs were unable to be scraped, due to errors generated in the session.
- Only one tale type per tale, intent is to store multiple ATUs as a nested list
- The `provenance` field is still messy, since multiple variables (i.e. country, region, tale collection) are still stored in a single column

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

6.1 Data Dictionary

The `aft` (i.e. *Annotated Folk Tales*) dataframe contains 904 rows, each corresponding to a single tale. Its 10 columns are described briefly below:

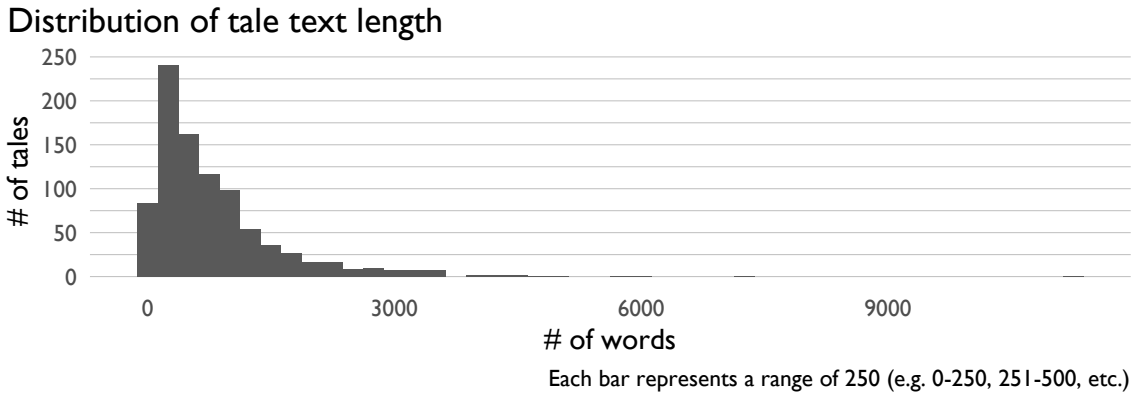
- `type_name` : The name associated with the Aarne-Thompson-Uther (ATU) tale type identifier.
- `atu_id` : The Aarne-Thompson-Uther (ATU) tale type identifier which classifies the tale.
- `tale_title` : The title of the tale.
- `provenance` : The person, place or tradition from which the tale came. In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

[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.

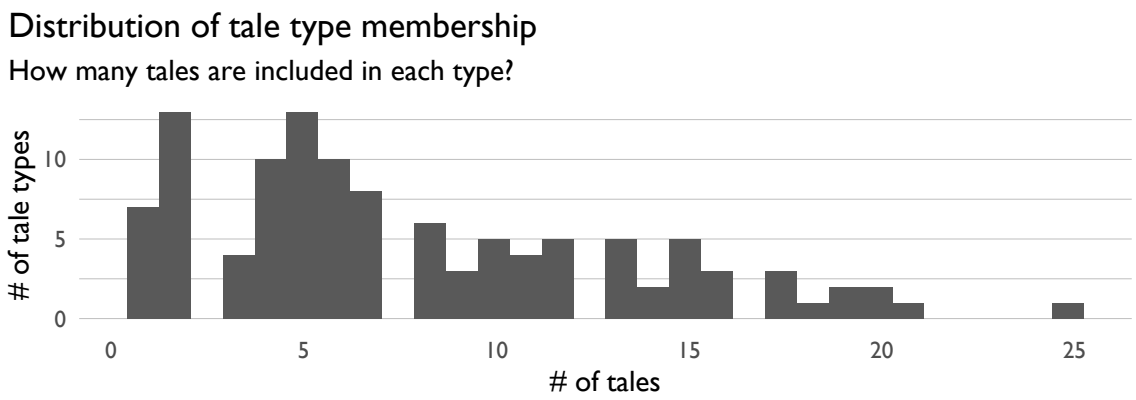
the tale came (*e.g. North Africa*), or the larger collection of tales in which the tale is found (*e.g. The Kathasaritsagara*).
- `notes` : Additional notes related to the tale.
- `source` : The bibliographic citation for the original published source of the tale.
- `copyright` : Any copyright information published alongside the tales in their scraped sources.
- `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.

6.2 Descriptive Statistics

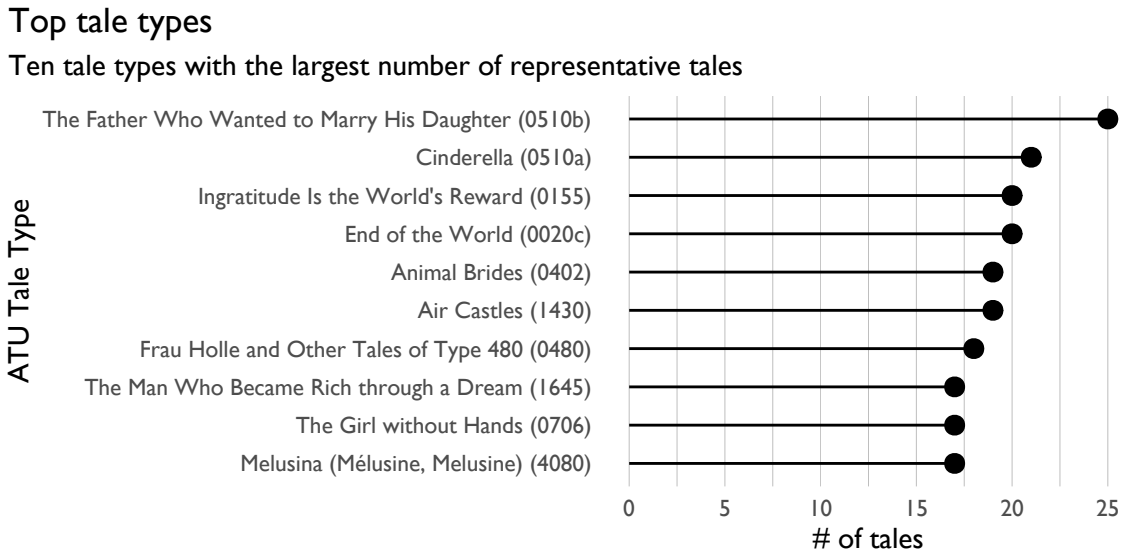
Length of tales. The 904 tales in the dataset average 833.1 words in length, though the individual texts vary with a minimum of 26 words and a maximum of 11210. The histogram below shows the distribution of tale lengths:



Number of tales by ATU type The tales compiled in the `aft` are annotated by Aarne-Thompson-Uther (ATU) tale type, and represent 113 distinct types. There are an average of 8 tales in each tale type, with a range of 1 to 25.



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

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 29:583–605, DOI 10.1093/llc/fqu055

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