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

Bag-of-tales

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 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 key · dictionary · word ·

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

We believe that a rendezvous between folk narrative studies and data science is already taking place on the Semantic Web as its venue. Limited in its scope for now, a next step in the evolution of Digital Humanities, this event has been long overdue, and will include more and more cultural artifacts of all ages and regions worldwide.

Regardless of how long it will take, and if one or more generations of multidisciplinary science will have to be called in, our professional vision and political stance is that there is one single body of facts about antiquity, including narratives in the realm of intangible cultural heritage. For common good, neither mankind nor the Semantic Web can afford to overlook this bulk of information about the past, independent of the attitude of funding agencies with other priorities. Therefore this single body of knowledge must be captured in its entirety and complexity. If a respective multidisciplinary top ontology will consist of a number of domain-specific knowledge graphs, only time will tell. But at this point in time and to this end, one needs a two-pronged strategy: the first effort is to convert the semantics inherent in those narratives to logical representations, the second one is to extract a complete set of facts by statistically robust relations. In other words, whichever way we want to proceed, research must address the relative lack of respective datasets.

We anticipate that, due to shortcomings on the funding side of the above equation, and to reduce any further delays, crowdsourcing of the problem via Github can be a legitimate approach, calling in data science for toolkit development. We will argue for this development below. When it comes to data science, missing folktale and myth datasets for machine learning, in English, and with semantic markup either by ATU or TMI tags, are a major bottleneck though. Likewise, for ontology building, ultimately one might need to convert the equivalents of a Roscher, a Pauly-Wissova, or respective monographs with specialist field knowledge over time. We [will] report related developments and considerations elsewhere (papers: Olympians, Burkert).

On the other hand, the methodology aspect of the above has seen progress. Recently, Yarlott & Finlayson have published a comprehensive overview of tale research (2016). Our vision subscribes to and includes their statements but notes the potential for more accomplishments. Further, we welcome and acknowledge results and proposals by (d'Huy, Tehrani, Berezhkin, Thangerlini, Meder, Karsdorp etc) who have been active in the dataset building and text analytics arena, or the ontology building efforts of (Declerck, Lendvai etc) as another research track. However, apparently these approaches and toolkits can be considerably extended as shown e.g. by Lendvai et al (Verona), and the direction we suggest points at the integration of more and more sophisticated NLP solutions, combined with evolving datasets in a data science framework, where the respective semantics is increasingly modelled by evolving ontologies and vector spaces (practically vector fields) vs. dynamic graphs, instead of static ones (D4.5, arXiv 1 & 2). Such developments will have to be extended to knowledge graphs as well. We also note in passing that, as suggested by

Ofek et al and Darányi et al, the analysis of tale types as motif strings in the framework of text variation invites the metaphor of narrative genomics (refs).

Our research problem for the current paper is as follows. Regarding folk narrative research, consider the case of two standard reference tools, the TMI and the ATU. The TMI has x_1 motifs (or, cf Yarlott & Finlayson, x_2), whereas the ATU uses y of them to model tale structures as motif strings on a global scale. It is justified to ask, where have w % of those motifs in the TMI disappeared by the time they were applied to the ATU; or, how can an important monograph acquire canonical status with such a discrepancy in its background? In our eyes, the explanation goes back to the very different comparison capacities of the human mind vs the computer, leading to differently robust deductions, and to remedy this situation is to call in data science.

The structure of this paper is as follows. In Section 2, we bring examples of research results relevant to our proposal, including a tentative overview of – mostly non-available – datasets in the field (Linked Open Data?). In Section 3, we discuss Ashliman's publicly available Folktexts dataset. In Section 4 data cleaning and data restructuring aspects are presented which lead to a dataset for machine learning. Section 5 lists a spectrum of first results, Section 6 ideas for future research. Section 7 is acknowledgements, Section 8 is the bibliography, followed by an appendix.

2 Relevant Related Research

Text with citations by [Galyardt(2014)].

3 The Ashliman Folktexts collection

as required. Don't forget to give each section and subsection a unique label (see Sect. 2).

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].

Strides have been made in the digital humanities to emulate these efforts, with the *Journal for Open Humanities Data* [cite] being a noteworthy exception to the more common practice.

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

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.
- *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 SHA to indicate the state of the dataset at the time the analysis was run, e.g. `https://github.com/j-hagedorn/trilogy/blob/f256a509633d06b206a58b5a21e5465b17d75`

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 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.

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

[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 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

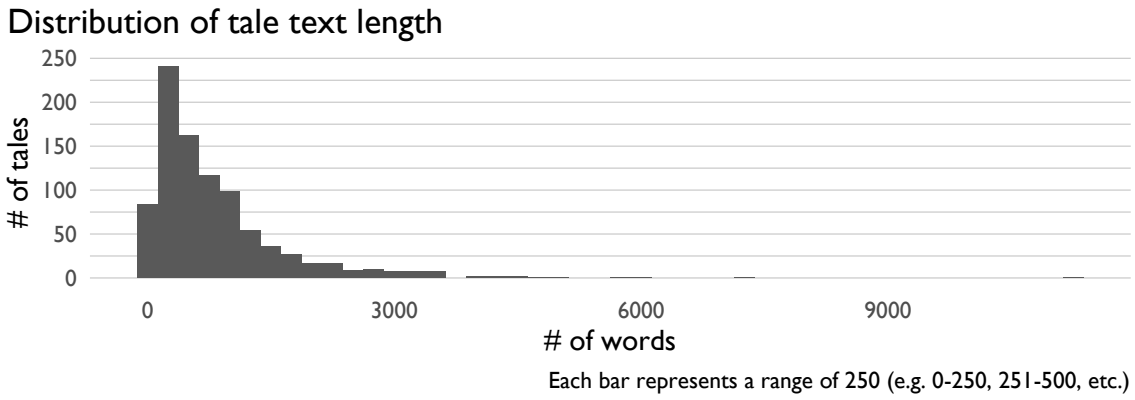
The `aft` (i.e. *Annotated Folk Tales*) dataframe contains 904 rows, each corresponding to a single tale. Its 10 columns are described briefly below:

- `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 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.

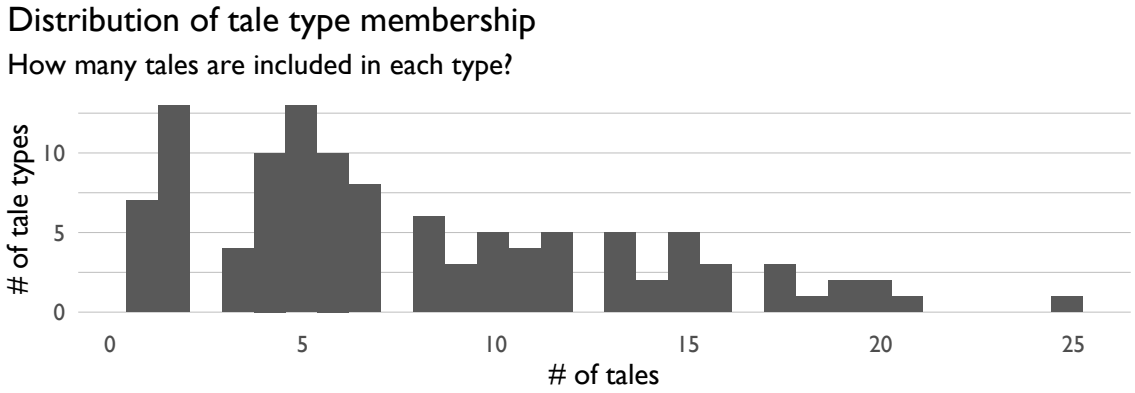
- `text` : The full text of the tale identified in `tale_title`.
- `data_source` : The source of the annotated tales. At the time of this writing, the source of all tales is "Ashliman's Folktexts", but this will change as the dataset grows.
- `date_obtained` : The date on which the data set identified as a `data_source` was last downloaded and compiled.

6.2 Descriptive Statistics

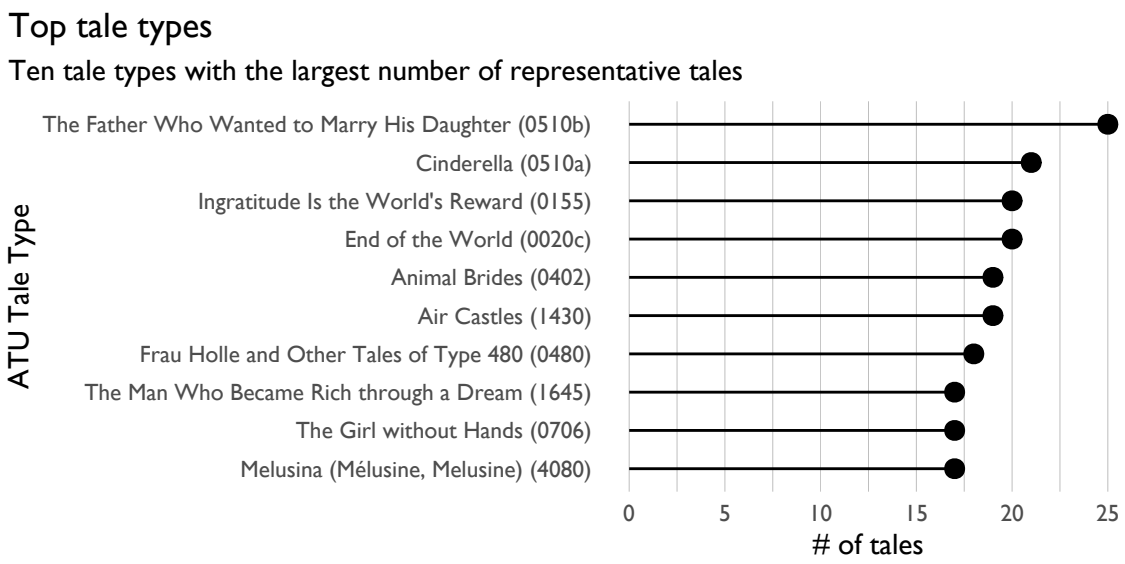
Length of tales. The 904 tales in the dataset average 833.1 words in length, though the individual texts vary with a minimum of 26 words and a maximum of 11210. The histogram below shows the distribution of tale lengths:



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

- TMI
- ATU

References

Galyardt(2014). Galyardt A (2014) Interpreting mixed membership models: Implications of erosheva's representation theorem. In: Airoldi EM, Blei D, Erosheva E, Fienberg SE (eds) Handbook of Mixed Membership Models, Chapman and Hall

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