What Makes a Fairytale

Five Factors of Fairytales

Jan Motl

University of Wisconsin-Madison
2924 Harvey Street 5H
Madison, Wisconsin, USA
jmotl@wisc.edu

Abstract

Traditionally fairytales were analyzed by their plot; however, this approach was criticized that it omits tone, mood, character and other things that further differentiates one fairy tale from another. To find characteristic moods of fairy tales, an approach used to get "Big Five" personality traits was applied to fairy tales. Adjectives from fairytales written in English were collected and analyzed using factor analysis. The analysis gave rise five unique factors describing moods of fairytales.

Folklorists have identified recurring pattern plots so that folklorists can organize, classify, and analyze the folktales they research. The first such classification was done by Aarne and later enlarged by Thompson (Aarne & Thompson, 1910) in *Verzeichnis der Märchentypen*. In the essay, "*The Motif-Index and the Tale Type Index; A Critique*," Alan Dundes explains that the Aarne-Thompson tale type index is one of the "most valuable tools in the professional folklorist's arsenal of aids for analysis" (Dundes, 1997). Similarly Russian folk tales were analyzed by Vladimir Propp in *Morphology of the Folk Tale* (Propp, 1928). In his work 31 different motifs were identified together with 8 characters. The advantage of Propp's classification is that it allows any combination of motifs and characters while Aarne –Thompson classification use categories and subcategories represented by a typical fairy tale. Thus a tail can't be both an animal tale and tale of the fantastic but only one of them. However Propp's work itself was criticized by Claude Lévi-Strauss for removing all verbal considerations from the analysis, even though the folktale's form is almost always oral, and also all considerations of tone, mood, character, and, anything that differentiates one fairy tale from another (Levi-Strauss, 1976). Nevertheless, these two systems are still in use (Tatar, 2003).

The purpose of this explanatory research is to find factors that describe moods of fairytales. A pioneer in the development of factor analysis L. L. Thurstone wrote in his report:

> Sixty adjectives hat are in common use or describing people … were given to each 1300 raters. Each rater was asked to think of a person whom he know well and to underline every adjective that he might may us in a conversational description of that person … the … correlation…coefficients for the sixty personality traits were then analyzed by means by means of multiple factor methods and we found that five factors are sufficient to account for the coefficients (Thurstone, 1934).

This work was later redone by Raymond B. Cattell, who began his personality explorations with a perusal of the approximately 4,500 trait-descriptive terms. And many other replicated the results with different terms but always with a variant of factor analysis (Goldberg, 1993). Hence it was decided to analyze adjectives in fairy tales with factor analysis.

Method

Fairytales

English translations of fairy tales were collected from two electronic databases: Project Gutenberg (Hart, 1971) and The Baldwin Project (Ripperton, 2000). Project Gutenberg is the largest collection of free electronic books; however, this collection mostly contains European literature and doesn't limit itself only on fairy tales. In contrast The Baldwin Project focuses on children's literature published in the USA. Unfortunately, no comparable electronic collections of African or Asian literature in English were found. Hence, most of the tales are of European heritage.

In total 1,989 fairy tales were collected amounting 4,058,101 words (20MB of text).

Data Preprocessing

In Big Five Factor analysis traditionally just the adjectives were used; however, fairy tales don't only describe nouns using adjectives (e.g. smart, handsome prince) but also verbs using adverbs (e.g. the prince quickly and smartly killed the dragon). Hence it was priory decided to work with frequency of both, adjectives and adverbs. The tales were tagged using OpenNLP (Baldridge, 2006) part of speech tagger and performance of the tagging was manually verified on a small sample of the tales. While the tagger tagged around 80% of adjectives and adverbs as adjectives or adverbs in the modern tales, it missed around 80% of adjectives and adverbs in the tales written in archaic English. Hence it was decided to try to use a list of the most common adjectives and adverbs as a look up list and see if it outperforms the tagger.

It was priory decided to use just a list of adjectives because many adverbs can be derived from adjectives (for example bluely from blue). The used list of 347 most common adjectives was obtained from Burst Media (Burst Media, 2011). Since adjectives can be inflected (for example adjective blue can take forms like bluish, bluer, bluest), the adjectives and all the words in the tales were reduced into their root forms' using Snowball stemmer. Consequently the frequency of these stemmed adjectives was calculated and stored into a matrix (fairytale versus stemmed adjective). When performance of this approach was manually checked on a small sample of the tales it was found that it reliably finds around 80% of adjectives and 40% of adverbs in both, new an old tales written in archaic English. The problem with this approach is that it identifies many other words as an adjective. For example adjective "informed" is stemmed into "inform" and this is a stem of both "information" and "inform" itself. Nevertheless, these words still carry some information about the mood of the fairy tale. Thus it was decided to use the simpler and faster method – look up list of adjectives.

Factor Analysis

Different method of analysis were tried ranging from clustering (k-means and hierarchical clustering with cosine similarity, which is commonly used in text mining, and expectation maximization clustering with Euclidean distance) over correlation and dependency similarity (including correlation and mutual information) to dimensionality reduction (principal component analysis, independent component analysis, singular value decomposition, self organizing maps and finally factor analysis). Even though mutual information, singular value decomposition and factor analysis were all giving meaningful categories of similar adjectives, it was factor analysis that gave the most easily interpretable categories.

Thus factor analyze, which finds similar words and put them under a common factor, is further discussed. The weighted least-squares estimate was used to predict factor scores and varimax setting was used to rotate the axis for easier interpretation of the results. The stability of the produced factors was verified by random dividing of the fairy tales into 5 groups. In each group the same interpretable factors emerged.

Study Results

Identified Factors

Factor analysis was run with different number of presumed factors for each group separately and the resulting boxplot of significances is in Figure 1.

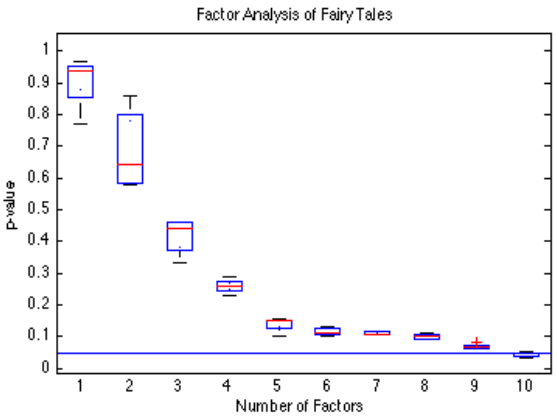

Figure 1: Boxplot of significance for 5 groups, where p-value is the right-tail significance level for the null hypothesis that the number of common factors is n. The blue horizontal line corresponds to significance level 0.05.

The significance level sharply decreases with each added factor till five factors are used. Then the significance level stays almost constant until nine or ten factors are used. Although results with at least ten factors were significant at confidence level of 0.05, it was decided to use five factors because the resulting plot bends at five factors; hence five factors has the best payoff between significance and simplicity. The significance was also calculated for factor analysis of the whole dataset and the resulting values were similar to the values with the groups. The calculated significance level for five factors was 0.1393. The five resulting factors for the first group are depicted in Table 1.

Table 1: First twenty stemmed adjectives sorted decreasingly (from top down) by the weight given by factor analysis. The results are for the first group of the fairy tales.

Factor I	Factor II	Factor III	Factor IV	Factor V
inform	delight	bright	cruel	adventur
necessari	splendid	green	hate	courag
accept	astonish	blue	tear	danger
frequent	magnific	color	wick	fierc
high	enchant	sweet	tender	huge
perfect	handsom	cold	ill	wander
ad	charm	fresh	griev	weari

obtain	spite	glorious	pain	faint
possess	sad	smile	sad	brave
agreeabl	guard	yellow	smile	victori
interest	imposs	tear	bitter	mighti
adventur	wick	pale	spite	mountain
entertain	marvel	gleam	gentl	swift
known	determin	sparkl	scary	fate
excel	ad	warm	sweet	thick
observ	hesit	clear	lone	thunder
success	bitter	long	faint	courag
remark	hideous	silent	wear	strong
famous	precious	fade	helpless	flung
natur	possess	plant	cold	sore

Interpretation

The interpretation of the factors is suggested in Table 2. The first factor was named Information because the first three terms (informed, necessary, and acceptable) are related to information or knowledge

The second factor was named Beauty because the first seven terms (delightful, splendid, astonishing…) are description of beauty. However, some terms like "sad", "wick" and "bitter" don't match this description and are more related to the fourth factor that was named Morbidity.

The third factor was named Color & Temperature because the first seven terms (bright, green, blue…) either describe color or temperature.

The fourth factor was named Morbidity because the first fourteen terms (cruel, hateful, tearful…) are related to morbidity. Out of all the factors this one is the most consistent and prevalent that I have to wonder whether fairy tales shouldn't be rather named gory tales. However, as Maria Tatar explains fairy tales without any cruelty are boring for both, children and adults. Hence even child fairy tales have to be spiced with some morbidity at least a bit.

The last factor was named Brave because the first eleven terms (adventurous, courageous, dangerous…) either describe a brave character or require a brave character.

Table 2: Interpretation of the factors. The terms in bold are considered to be related to the assessed name of the category. Antonyms are highlighted as well when it is appropriate.

Information	Beauty	Color & Temp.	Morbidt.	Brave
inform	**delight**	**bright**	**cruel**	**adventur**
necessari	**splendid**	**green**	**hate**	**courag**
accept	**astonish**	**blue**	**tear**	**danger**
frequent	**magnific**	**color**	**wick**	**fierc**
high	**enchant**	**sweet**	**tender**	**huge**
perfect	**handsom**	**cold**	**ill**	**wander**
ad	**charm**	**fresh**	**griev**	**weari**
obtain	spite	glorious	**pain**	**faint**
possess	sad	smile	**sad**	**brave**
agreeabl	guard	**yellow**	**smile**	**victori**
interest	imposs	tear	**bitter**	**mighti**
adventur	wick	**pale**	**spite**	mountain
entertain	**marvel**	**gleam**	**gentl**	**swift**
known	determin	**sparkl**	**scary**	fate
excel	ad	**warm**	sweet	thick
observ	hesit	**clear**	**lone**	thunder
success	bitter	long	**faint**	**courag**
remark	**hideous**	silent	**wear**	**strong**

Usability

It was found desirable to evaluate how well these factors reflect the mood of the tales. Hence for each factor a story with the highest score in that factor was inspected.

The Information factor is highest in "The Story of Prince Ahmed and The Fairy Paribanou: *"…as he was obliged to stay there for his brothers as they had agreed, and as he was curious to see the King of Bisnagar and his Court, and to inform himself of the strength, laws, customs, and religion of the kingdom…"* These stories tend to be spoken in formal manner.

The beauty factor starts the plot in "The Son of Seven Queens": *"…Her beauty bewitched him, so he fell on his knees, begging her to return with him as his bride…"*

Color & Temperature is represented by "The Little Mermaid". This fairy tale is full of adjectives, as it begins with: "*Far out in the ocean, where the water is as blue as the prettiest cornflower, and as clear as crystal, it is very, very deep, so deep…"*

And "The Babes in the Wood", an example of Morbidity, pretty quickly turns morbid: "*…By the end of this time the gentleman fell sick, and day after day he grew worse. His lady was so much grieved by his illness that she fell sick too. No physic, nor anything else, was of the least use to them, for they grew worse and worse...*"

The Brave factor is the highest in "St. George of Merrie England": "*…Now, when twice seven years had passed the boy began to thirst for honorable adventures, though the wicked enchantress wished to keep him as her own. But he, seeking glory…*"

The list of the tales that scored highest in each respectable factor is in Figure 2.

Information
'Story of Prince Ahmed and Fairy Paribanou'
'The Story of the Fisherman and the Genie'
'Wizard King'

Beauty
'The Son of Seven Queens'
'The Golden Branch'
'The White Cat'

Color & temperature
'The Little Mermaid'
'Snow Queen'
'The Story of the Year'

Morbidity
'The Babes in the Wood'
'The Nettle Spinner'
'White Doe'

Brave
'St. George of Merrie England'
'The Enchanted Pig'
'Lady of Fountain'

Figure 2: List of the tales that scored highest on each respectable factor

Validity

It is necessary to realize that the analyzed dataset is very limited in the size and diversity of the sources. For example over 10% of all fairytales (213 from 1989) are from Brothers Grimm. Hence the found factors can well describe the fairytales in the dataset but completely fail in the real world. Therefore some tales, which were not present in the training dataset, were analyzed. These tales are presented in Table 3 and although they all are present at Project Gutenberg, they have been omitted from the training set because they haven't been properly categorized as fairy tale. Hence they have been overlooked.

To be able sensibly compare fairy tales among themselves the factor scores of fairy tales in the training set have been normalized to the same mean and variance and subsequently mapped between 0 and 100, where 100 means that the tail scores on the factor as much as the "exemplary" tail in the training set. Unfortunately, some of the tails are so short that they score 0 on all the factors because they don't use any word from the look up list. This is almost a case of "The Pied Piper of Hamelin" because only a rhymed version of the tail is on Project Gutenberg and is quite short.

Table 3: Score of selected fairy tales on the five factors.

	I.	**II.**	**III.**	**IV.**	**V.**
Alice in Wonderland	42	30	85	73	62
Peter Pan	56	71	87	67	43
Pinocchio	22	13	58	59	27
The Pied Piper of Hamelin	0	0	36	35	14

Discussion

A problem rise when we look at the selection of the words of interests. What would we get if adverbs, and not adjectives, were used in the look up list? Or what if different adjectives were used? Unfortunately the answers for these questions are left on follow up work.

Other unanswered question is what happens if a specific subgroup of fairy tales is analyzed. Would the results look different if only Celtic tales were analyzed? If only Grimm's tales were analyzed? The probable answer is yes because during the cluster analysis Celtic tales often created a small but well separated cluster. And each author use different style and words. This feature was for example used in the analysis of Bible authorship (Friedman, 2011).

Another issue is that the dividing of adjectives into five factors is not significant. The source data are freely available at http://motl.us/wiki/doku.php/public/fairytales for further analysis.

Conclusions

The objective of the study was to find factors describing mood in the fairy tales. The study found five factors: Information, Beauty, Color & Temperature, Morbidity and Brave. The study did for fairytales the same big five factors did for human personality.

References

Aarne, A., & Thompson, S. (1910). *Verzeichnis der märchentypen.* Berlin.

Baldridge, J. (2006). *OpenNLP*. Retrieved from http://opennlp.sourceforge.net/projects.html

Burst Media. (2011). *List of Adjectives.* Retrieved from http://www.momswhothink.com/reading/list-of-adjectives.html

Dundes, A. (1997). The Motif-Index and the Tale Type Index; A Critique,. *Journal of Folklore Research*, 195.

Friedman, M. (2011). *Huffpost.* Retrieved from http://www.huffingtonpost.com/2011/06/29/an-israeli-algorithm-shed_n_886996.html

Goldberg, L. R. (1993). The Structure of Phenotypic Personality Traits. *American Psychologist*, 26-34.

Hart, M. (1971). *Project Gutenberg.* Retrieved from http://www.gutenberg.org/

Levi-Strauss, C. (1976). *Structure and Form: Reflection on a Work by Vladimir Propp.* London: Allen Lane: Vol. 2. Trans. Monique Layton.

Propp, V. Y. (1928). *Morphology of the Folktale.* 2nd Ed.University of Texas Press.

Ripperton, L. (2000). *The Baldwin Online Children's Literature Project.* Retrieved from http://www.mainlesson.com

Tatar, M. (2003). *The Hard Facts of the Grimms' Fairy Tales.* Princeton: Princeton University Press.

Thurstone, L. L. (1934). The vectors of mind. *Psychological Review*, 12-14.