PATTERN RECOGNITION OF CHILDREN STORIES BASED ON THEMES USING TIME SERIES DATA

Ms. Menaka Sikdar[1] and Ms. Pranita Sarma[2]

[1]Ph.D Research scholar, Department of Statistics, Gauhati University, Guwahati, Assam, India
[2]Professor, Department of Statistics,Gauhati University, Guwahati, Assam, India.

ABSTRACT

This paper presents a study for three languages namely Assamese, Bengali and English. The main objective of this study is to recognize the patterns based on language model with special reference to children stories in order to find the distinction among all these languages. We consider only the children stories because they are found to be similar all over the world with different flavors produced by different cultures, languages and time. Sixteen stories based on themes namely Fairy tales, Fables, Jack tales and Formula tales are considered for analysis. The significant differences among the types of stories written by different authors are verified . Autocorrelation test is performed by using Box- Ljung statistic to determine the randomness of the data by observing the language data as time series data. However for some of the stories, the data are found to be random and hence Kolmogorov goodness of fit test and Smirnov test are conducted to test the significant differences only for those stories. It has been shown that there exist significant differences among the writing patterns of the children stories written by different authors in different languages.

KEYWORDS

Autocorrelation, Box- Ljung statistic, Empirical distribution, Kolmogorov Goodness –of- Fit test and Smirnov test.

1. INTRODUCTION

The maximum numbers of children stories are compiled from our classical folklores and folktales. Therefore the structures of these stories may be changed but the skeleton of the stories along with the moral lessons that are intended to inject the society remain all the more same over time and places. Conventional Indo- European stories may be structurally different in many subjects but their bases and themes are same. The variation in the writing patterns of stories with respect to different languages, cultures and authors seems to quite natural. The skeleton of the stories being same, it is difficult to comment whether there exists any significant differences amongst them at their first look. The children stories are mainly classified into four different categories namely Fairy tales[1], Fables[2], , Jack tales[3] and formula tales[4] . Understanding of the complexity associated with the pattern of writing offered by different authors in various languages is not an easy task. All kinds of mathematical tools were adopted to gain understanding the complexity of human language. One of these tools is Time Series Analysis (D.S.G Pollock, 1999). Time series analysis plays a key role not only in physical sciences but also in statistical sciences. Time series analysis has been used to investigate written human language texts by Kosmas Kosmidis et al(2006) and DENG W.B.et al(2011). Sikdar and Sarmah (2017A,2017B,2017C and 2017D) presented four articles on three languages namely Assamese, Bengali and English for pattern recognition of language model with special reference to children stories of which one is based on direct speech. In this article sixteen stories based

DOI: 10.5121/ijnlc.2018.7504

on themes namely fairy tales, Fables, Jack tales and Formula tales are considered for analysis. Autocorrelation test is performed by using Box- Ljung statistic to determine the randomness of the data by observing the language data as time series data. However for some of the stories, the data are found to be random and hence Kolmogorov goodness of fit test and Smirnov test are conducted to test the significant differences only for those stories. The significant differences among the types of stories written by different authors are verified . Section 2 of this article is about the objective of the study whereas the source of data is presented in section 3 . The materials and methods used to perform the analysis are given in section 4. Section 5 contains the results of analysis.

2. OBJECTIVE OF THE STUDY

The main objectives of our study are

a) To gain understanding of the process generating the sentence length time series (SLTS) [mentioned in section 4.1]
b) To recognize the pattern of writings presented by different authors in three different languages namely Assamese, Bengali and English with respect to the types of stories.
c) To study the significance differences among the stories which are theme wise similar.

3. SOURCES OF DATA

For analyzing the children stories based on themes, the data have been collected from the stories by Sahityo –rothi Laxminath Bezbarua**,** Upendrakishore Roy Choudhury and the Grimm brothers.The stories from each category(Fairy tales, Fables, Jack tales and Formula tales) which are theme wise similar in 'Burhi Aai'r Xaadhu'(Assamese) ,'Tuntunir Boi' (Bengali) and Grimm's Fairy Tales (English) are selected for the purpose of analysis. The number of words in a sentence and number of sentences of the corpus under consideration are counted by using Microsoft Office Excel 2007. The lists of theme wise similar stories are given in table1.

Table 1.List of selected stories

Category	Similar Stories
Fairy tales	(i)Rapunzel (R) [English] (ii) Silonir jiyekar Xaadhu(SJX)[Assamese]
Fables	(iii)Budhiyak Sheal (BS)[Assamese] (iv) Majantali Sarkar (MS) [Bengali]
	(v)Bandar aaru Sheal (BAS)[Assamese] (vi) Cat and Mouse in partnership (CAMIP)[English]
	(vii)Charai aar bhager katha (CABK) [Bengali] (viii) The Dog and the sparrow(TDATS) [English]
Jack Tales	(ix) Sarabjaan (Sj) [Assamese] (x) Doctor knowall (DK)[English]
	(xi) Buraa Burhi aaru Sheal(BBAS) [Assamese] (xii) Kujo burhir Katha(KBK) [Bengali]
	(xiii)Panta-Burhir katha (PBK) [Bengali] (xiv) Clever Gretel(CG) [English]
Formula Tales	(xv) Dorakauri aaruTipchicharai (DAT) [Assamese] (xvi) Charai aar kaker katha (CAKK) [Bengali]

4. MATERIAL AND METHODS

To achieve the objectives of our study, we have performed Autocorrelation test using Box- Ljung statistic to determine the randomness of the data. However for some of the stories, the data are found to be random and hence Kolmogorov goodness of fit test, Smirnov test, Kruskal-Wallis test

and Squared rank test (Conover W.J. (2006)) have been conducted to test the significant differences among the stories.

4.1. Language Time Series

At a glance, it does not seem that language and time series have any relation between them. Before applying the time series analysis method, we can observe the written document (story) as time series data. Let us take a document (story) having 'n' sentences and S_i (i=1,2,…,n) be the length (total number of words) of the i^{th} sentence .The position of the sentence in a written document depends on time. Let t_i be time epoch at which the completion of i^{th} sentence takes place $\forall\, i = 1,2,\dots n$. Hence our data set contain $(t_i, S_i)\ \forall\, i = 1,2,\dots n$.

4.2. Autocorrelation Test

Autocorrelation test has been performed to detect the randomness of the distribution of the sentence length of different stories written in three different languages namely Assamese, Bengali and English. The autocorrelation function [ACF] can be used to detect non-randomness in data and also to identify an appropriate time series model if the data are not random. For our data set (S_i,t_i), i=1,2,…,n, the lag k autocorrelation function may be defined as is

$$r_k = \frac{\sum_{i=1}^{n-k}(S_i - \bar{S})(S_{i+k} - \bar{S})}{\sum_{i=1}^{n}(S_i - \bar{S})^2}$$

The value of r_k ranges from -1to 1, and the larger the absolute value of r_k, the stronger the correlation in the time series. The randomness is ascertained by computing autocorrelation for data set at varying time lags. If random such autocorrelations should be near zero for any and all time-lag separations. If non-random then one or more of the autocorrelations will be significantly non zero. On the other hand, the Box- Ljung test (1978) is used to test whether or not observations over time are random and independent. In particular, for a given *k*, it tests the following hypotheses:

H_{0j}: The autocorrelations of the length of sentences up to lag *k* under jth story are all zero for

j=1,2,…,m

H_{1j}: The autocorrelations of the length of sentences of jth story at one or more lags differ from

0.

The test statistic is calculated as:

$$Q_k = n(n+2)\sum_{t=1}^{k}\frac{r_t^2}{n-t}$$

which approximately follows chi-square distribution with k degrees of freedom.

4.3. The Empirical Distribution Function

The concept of Empirical distribution function is used for studying the probabilistic structure of the distributions of sentence lengths of various stories which are randomly distributed. It is interesting to note that except formula tales randomness is exhibited in other categories of the stories. In case of **the distribution of length (number of words) of sentences** of a

particular story, our data consist of a random sample $S_1, S_2,\ldots,S_n$ **of size n.** The empirical distribution function,

$$F_s(x) = (\text{number of } s \leq x)/n.$$

4.4. Kolmogorov Goodness –Of- Fit Test (One Sample Kolmogorov And Smirnov Test)

Our objective is to verify whether these random variables are normally distributed. The Kolmogorov Goodness –of- Fit test is used for this purpose. **The distribution of length (number of words) of sentences** of a particular story ,constitutes a random sample $S_1,S_2,\ldots,S_n$ **of size n** associated with some unknown distribution function , denoted by Q(s).

Test Statistic

Let F(s) be the empirical distribution function based on the random sample $S_1,S_2,\ldots,S_n$ of size n. Let $Q^*(s)$ be a completely specified hypothesized distribution function which is considered here as a normal probability distribution function.

(Two- Sided Test) Let the test statistic T_1 be the greatest vertical distance between F(s) and $Q^*(s)$. Mathematically $T_1 = \sup_s |Q^*(s) - F(s)|$

Null Distribution: when Q(s) is continuous and the null hypothesis is true, the approximate distribution function of T_1 is $p(T_1 \leq s) = [G(s)]^2$

where

$$G(s) = 1 - s \sum_{p=0}^{[n(1-s)]} \binom{n}{p} \left(1 - s - \frac{p}{n}\right)^{n-p} \left(s + \frac{p}{n}\right)^{p-1}$$

Where *[n (1-s)]* is the greatest integer less than or equal to *n(1-s).*

Hypotheses

The null hypothesis is to be tested

H_0: $Q(s)=Q^*(s)$ for all s from $-\infty$ to $+\infty$
H_1: $Q(s)\neq Q^*(s)$ for at least one value of s

4.4. Smirnov Test (Also Known As Two Sample Kolmogorov And Smirnov Test)

Kolmogorov and Smirnov developed statistical procedure that uses the maximum vertical distance between two empirical distribution functions as a measure of how well the functions resemble each other. Our objective is to compare the distributions of two stories which are theme wise similar but written by different authors in different languages . The lists of such similar stories given in table 1 are being considered for our analysis. Here we are trying to determine whether the two distribution functions of the length of sentences under two similar stories are identical by applying Smirnov test.

Let $S_{11}, S_{12},\ldots,S_{1n_1}$ be the length of $\mathbf{n_1}$ **sentences under story 1** associated with some unknown distribution function , denoted by $Q_1(s)$ and $S_{21}, S_{22},\ldots,S_{2n_2}$ be the length of $\mathbf{n_2}$ **sentences under**

story 2 associated with some unknown distribution function , denoted by $Q_2(s)$.

Test Statistic

Let $F_1(s)$ and $F_2(s)$ be the empirical distribution function of the length of sentences under story1 and story2 respectively. The test statistic for two-sided test is defined as

$$T_2 = \sup_s |F_1(s) - F_2(s)|$$

Null Distribution

The exact null distribution of T_2 is obtained by considering all ordering of S_1's and S_2's to be equally likely under the null hypothesis and computing T_2, as appropriate, for each ordering .Quantiles of the null distribution are given in table 19 for equal sample sizes and table 20 for unequal sample sizes. [Table 19 and Table 20 are given in "Conover W.J. (2006)]

Hypothesis: (Two sided test)

H_0: $Q_1(s) = Q_2(s)$ for all s from $-\infty$ to $+\infty$

H_1: $Q_1(s) \neq Q_2(s)$ for at least one value of s

5. ANALYSIS OF DATA

Autocorrelations and results of Box- Ljung test of Sentence length time series of different stories under different languages are obtained by using SPSS[5] software and are given in table 2.

Table2. Results of autocorrelation and Box-Ljung test

Language	category	story	lag	Autocorrelation	Box-Ljung statistic		
					value	d.f	p-value
English	Fairy tales	R	1	-0.044	0.112	1	0.74
			2	-0.044	0.227	2	0.90
			3	-0.208	2.826	3	0.42
			4	-0.235	2.900	4	0.58
			5	0.245	6.673	5	0.25
Assamese	Fairy tales	SJX	1	0.183	4.602	1	0.03
			2	0.083	5.558	2	0.06
			3	-0.030	5.685	3	0.13
			4	0.018	5.731	4	0.22
			5	-0.065	6.329	5	0.28
Assamese	Fable	BS	1	0.140	1.402	1	0.24
			2	0.002	1.402	2	0.50
			3	0.027	1.455	3	0.69
			4	0.104	2.256	4	0.69
			5	0.117	3.284	5	0.66
Bengali	Fable	MS	1	0.143	2.380	1	0.12
			2	0.116	3.963	2	0.14
			3	-0.023	4.025	3	0.26
			4	-0.201	8.827	4	0.07
			5	-0.022	8.886	5	0.15
Assamese	Fable	BAS	1	0.218	3.410	1	0.07
			2	0.292	9.650	2	0.008*
			3	0.077	10.091	3	0.018
			4	0.088	10.676	4	0.030
			5	0.134	12.058	5	0.034

English	Fable	CAMIP	1	0.173	1.650	1	0.199
			2	0.106	2.276	2	0.320
			3	-0.064	2.510	3	0.474
			4	-0.122	3.386	4	0.495
			5	0.042	3.490	5	0.625
Bengali	Fable	CABK	1	0.297	4.696	1	0.030
			2	0.240	7.826	2	0.020
			3	-0.039	7.909	3	0.048
			4	-0.085	8.320	4	0.081
			5	-0.067	8.580	5	0.127
English	Fable	TDATS	1	-0.085	0.449	1	0.503
			2	-0.231	3.821	2	0.148
			3	0.157	5.414	3	0.144
			4	0.201	8.045	4	0.090
			5	-0.094	8.634	5	0.125
Assamese	Jack tale	Sj	1	0.046	0.189	1	0.664
			2	0.073	0.686	2	0.710
			3	0.044	0.864	3	0.834
			4	0.004	0.865	4	0.930
			5	0.147	2.917	5	0.713
English	Jack tale	DK	1	-0.148	0.837	1	0.360
			2	-0.093	1.179	2	0.555
			3	-0.089	1.499	3	0.683
			4	0.061	1.655	4	0.799
			5	0.009	1.659	5	0.894
Assamese	Jack tale	BBAS	1	0.098	0.477	1	0.490
			2	-0.175	2.046	2	0.359
			3	-0.289	6.429	3	0.093
			4	-0.212	8.836	4	0.065
			5	-0.125	9.691	5	0.084
Bengali	Jack tale	KBK	1	0.032	0.078	1	0.780
			2	0.064	0.400	2	0.819
			3	-0.087	0.994	3	0.803
			4	0.033	1.079	4	0.898
			5	-0.058	1.356	5	0.929
Bengali	Jack tale	PBK	1	-0.007	0.003	1	0.957
			2	0.002	0.003	2	0.998
			3	0.018	0.022	3	0.999
			4	0.048	0.164	4	0.997
			5	-0.010	0.171	5	0.999
English	Jack tale	CG	1	-0.199	1.738	1	0.187
			2	0.212	3.773	2	0.152
			3	-0.292	7.716	3	0.052
			4	0.045	7.812	4	0.099
			5	-0.156	9.010	5	0.109
Assamese	Formula tales	DAT	1	0.260	5.830	1	0.016
			2	0.123	7.155	2	0.028
			3	-0.204	10.84	3	0.013
			4	-0.251	16.47	4	0.002*
			5	-0.111	17.59	5	0.004*
Bengali	Formula tales	CAKK	1	-0.184	1.697	1	0.193
			2	-0.157	2.965	2	0.227
			3	0.652	25.21	3	0.000*
			4	-0.155	26.50	4	0.000*
			5	-0.007	26.50	5	0.000*

*[The p-values with * marks indicate the rejection of the null hypothesis both at 5% and 1% level of significance (that indicates non-randomness). Otherwise the null hypotheses are accepted (that indicates randomness)]*

It is interesting to note that the autocorrelation between the lengths of the sentences does not have any particular pattern. At certain lag they are found to be very small whereas sometime their absolute values are found to slightly bigger than 0.2. For this purpose we have to use Box- Ljung statistic to verify the randomness of the data. Three categories namely Fairy tales, Fables and jack tales are found to be random in nature whereas formula tales show non randomness in the data under consideration. As mentioned in the note under table 2, the randomness of the length of the sentences of the stories under different categories may be decided. **The randomness of the data are exhibited for stories of all categories except for formula tales**.

The empirical distribution functions have been plotted for similar (theme wise) stories in the same graph by using R^6 Language and they are given below

[Assamese→ ,Bengali→ ,English→]

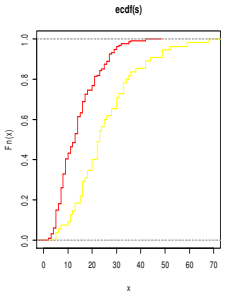

Figure1.(R & SJX)

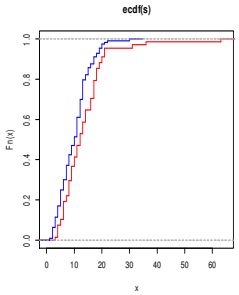

Figure2.(BS & MS)

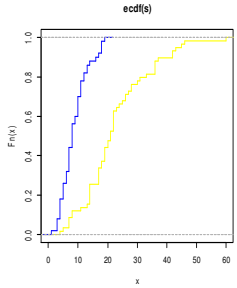

Figure3.(TDATS & CABK)

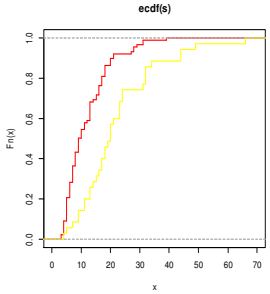

Figure4.(Sj & DK)

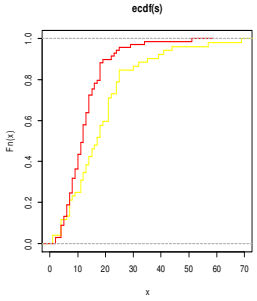

Figure5.(BAS & CAMIP)

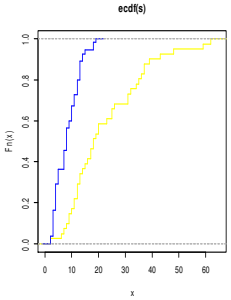

Figure6.(PBK & CG)

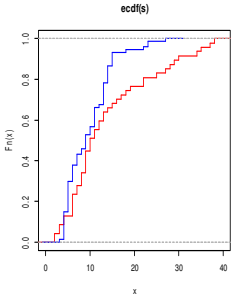

Figure7.(BBAS & KBK)

[The empirical distribution functions of length of sentences of different stories under Assamese, Bengali, English are represented in Figures 1,2,3,4,5,6 and 7 respectively]

Results of Kolmogorov Goodness- Of-Fit Test are obtained by using **SPSS soft-ware** and are given in table3

Table3. Results of One-Sample Kolmogorov-Smirnov Test for Normality

Language	**Story**	**Sample size(n)**	**Mean (μ)**	**S.D (σ)**	**Absolute difference (D)**	**K.S.Test statistic (Z)**	**p value**
English	R	55	25.51	13.60	0.119	0.880	**0.421**
	CAMIP	52	18.75	13.42	0.167	1.204	**0.110**
	TDATS	59	22.81	11.48	0.155	1.193	**0.166**
	DK	35	22.23	13.28	0.190	1.123	**0.160**
	CG	41	22.73	14.25	0.161	1.033	**0.236**
Assamese	SJX	134	14.05	8.35	0.130	1.509	**0.021**
	BS	68	13.28	8.738	0.147	1.216	**0.104**
	BAS	69	12.61	7.81	0.154	1.279	**0.076**
	Sj	88	11.85	7.46	0.149	1.398	**0.040**
	BBAS	47	14.02	9.95	0.179	1.228	**0.098**
Bengali	MS	113	9.97	5.34	0.083	0.879	**0.422**
	CABK	50	8.8	4.45	0.131	0.929	***0.354***
	KBK	74	9.70	5.18	0.141	1.212	***0.106***
	PBK	55	8.33	4.45	0.136	1.010	***0.259***

From table 3, it has been noticed that the p-values of the test statistics for **the distributions of length of sentences** of different stories(**except for SJX and Sj**) under different languages are greater than 0.05.Therefore we may accept our null hypotheses at 5% level of significance . On the other hand the p-values of the test statistics under the stories namely SJX and Sj(Assamese) are greater than 0.01 and hence the null hypotheses of these stories may be accepted at 1% level of significance. Hence we may conclude that the distributions of length of sentences of different stories under different languages namely Assamese, Bengali and English are **normally distributed**.

Results of **Smirnov Test** are obtained by using **SPSS soft-ware** and are given in table4

Table4. Results of Two-Sample Kolmogorov-Smirnov Test

Stories	**Language**	***Category***	***Sample size(n)***	***Absolute difference (D)***	***K.S. Z***	***p value***
R	English	Fairy	55	0.430	2. 686	0.000
SJX	Assamese		134			
BS	Assamese	Fable	68	0.211	1.377	0.045
MS	Bengali		113			
TDATS	English	Fable	59	0.707	3.680	0.000
CABK	Bengali		50			
BAS	Assamese	Fable	69	0.307	1.672	0.007
CAMIP	English		52			
Sj	Assamese	Jack	88	0.425	2.215	0.000
DK	English		35			
BBAS	Assamese	Jack	47	0.252	1.349	0.053
KBK	Bengali		74			
CG	English	Jack	41	0.561	2.721	0.000
PBK	Bengali		55			

From table 4, it has been noticed that the p-values of the test statistics for **the distributions of length of sentences** of similar stories under English and Assamese as well as Bengali and English are less than 0.05 and we may reject our null hypotheses at 5% level of significance. Therefore we may conclude that the distributions of length of sentences of different stories under English are significantly different from Assamese and Bengali. On the other hand, the p-value of the test statistic for the stories namely BS (Assamese) and MS(Bengali) is greater than 0.01 and the null hypothesis may accepted at 1% level of significance. Again the p-value of the test statistic for the stories namely BBAS (Assamese) and KBK (Bengali) is greater than 0.05 and the null hypothesis may accepted at 5% level of significance. Therefore we may conclude that the distributions of length of sentences of Assamese and Bengali stories are not significantly different.

6. CONCLUSION

This study presents the variation of children stories with respect to language ,similar themes and authors. The stories are written in the period of 19^{th} and 20^{th} centuries and we observed similarity and also dissimilarity in many cases. Autocorrelation test shows that the sentence structures only for formula tales are found to be non-random. However while analyzing the children stories in all three languages; it is observed that there is a striking dissimilarity between modern Indian languages and English. The sentence structures of English stories are significantly different from Assamese and Bengali stories. It has been noticed that pattern corresponding to Assamese and Bengali stories are similar to a great extent . One of the possible reasons is that they are originating from same Sanskrit language and hence further investigation corresponding to grammatical structure is necessary. Our future work will be devoted to pattern recognition based on grammar.

REFERENCES

[1] Amit M. Shemerler Y. and Eisenberg E,(1994) “Language and codification dependence of long range correlation in texts,” Fractals,vol. 2,pp7–13

[2] Ausloos M. (2008) “Equilibrium and dynamic methods when comparing an English text and its Esperanto translation” Physica A, vol. 387,pp 6411–6420.

[3] Conover W.J. (2006) Practical Nonparametric Statistics (3rd ed.), John Willey and Sons Inc.

[4] Deng W B. Wang D J. Li W. et al. (2011) “English and Chinese language frequency time series Analysis” Chinese Science Bulletin, vol. 56, No.34, pp 3717– 3722.

[5] Duda Richard O., Hart Peter E. Stork David G.(2000). Pattern Classification (2nd ed.), John Willey and Sons Inc.

[6] Gujarati D.N., Porter D.C. & Gunasekar S.,(2012) Basic Econometrics(5th ed), McGraw Hill Education (India) Private Limited

[7] Kosmidis K, Kalampokis A, Argyrakis P. (2006) “Language time series analysis” Physica A, vol. 370 ,pp 808–816.

[8] Ljung G M and Box G E P.(1978) “On a measure of a lack of fit in time series models”Biometrika, vol. 65, pp 297–303.

[9] Maurice G. Kendall and William R. Buckland, (1971) A Dictionary of Statistical Terms, Hafner Publishing Company, New York.

[10] Mukhopadhyay P., (2005) Applied Statistics, (2nd ed.).Books and Allied (P) Ltd

[11] Schenkel A, Zhang J, Zhang Y C.(1993) "Long range correlation in human writings" Fractals, vol. 1, pp 47–57

[12] Sikdar M. & Sarmah P., (2017A) "Pattern Recognition In Language Model With Special Reference To Children Stories" International Journal of Innovative Research and Advanced Studies (IJIRAS), Vol. 4, Issue3,pp 397-406.

[13] Sikdar M. & Sarmah P., (2017B) "Statistical Pattern Classification Of Direct Speeches In Children Stories" International Journal of Applied Mathematics and Statistical Sciences, ISSN(P): 2319-3972; ISSN (E): 2319-3980, Vol.6, Issue 5, pp. 7-18.

[14] Sikdar M. & Sarmah P. (2017C)."Pattern Recognition Of Children Stories With Special Reference To 'Jack Tales'" International Journal of Applied Mathematics and Statistical Sciences (IJAMSS), ISSN (P): 2319-3972; ISSN (E): 2319-3980, Vol.6, Issue 5; 81-90.

---

[1]Fairy tale-: A fairy tale is usually about a magic. There may be a witch but good always triumphs over bad. Often the characters are children, who beat the evil person and 'live happily ever after'.
[2]Fable-: It uses animals, legendary creatures, plants, inanimate objects, or natures that are anthropomorphized to explain moral values to the children.
[3]Jack tale-: A Jack tale is a category of 'folk tale', originating in the Middle Ages, they usually have a character called Jack who appears lazy or stupid but actually wins in the end because he is 'tricky.
[4]Formula tale: A formula tale comes under a category where a chain is created in the story that goes on increasing in size towards the end of the story.
[5]*www.ibm.com/analytics/us/en/technology/spss*
[6]*www.r-project.org*

AUTHORS

Ms Menaka Sikdar She is a research scholar in the department of Statistics, Gauhati University. She is pursuing her research work in "language pattern Classification".

Ms Pranita Sarmah She is a retired professor in the department of Statistics, Gauhati University. Her research interests are Reliability theory, Statistical Inference and Stochastic modeling. She has several publications in the areas of health, finance and various social issues. Recently she has published some papers in "Stochastic modeling in Indian classical Music" .Presently, She is working in "language pattern Classification".