FOLK TALES PLOTS CLUSTERING

Serge Potemkin (prolexprim@gmail.com) Lomonosov MSU,Moscow, Russia

Keywords: plot, folklore, distance, lexics, scaling, cluster.

Abstract

This paper deals with the automatic calculation of lexical similarity measures between the plots of Russian fairy tales. We use the Comparative index of plots which was stored in the relational database. Firstly we consider the "bag of words" method which turned out to be insufficient, and then we develop a method for word sequence to word sequence matching in two plots. This method occurred to produce better results. Expert evaluation of the plot pairs from the top of the list of the tales proximity shows considerable matching between plot pairs. Then the multidimensional scaling of all neighboring plots onto 2-dimension space made it possible to display the diagrams for visual investigation. One can browse clusters with different degrees of details for further evaluation of the actual similarities / differences of the plots. The obtained tool could be used for folklore studies, as well as for the construction of thesauri in other subject areas, in particular it was used for English and Nepalese tales.

Keywords: plots index, dynamic programming, multidimensional scaling.

КЛАСТЕРИЗАЦИЯ СКАЗОЧНЫХ СЮЖЕТОВ

Introduction

Structuring the vast material collected by folklorists during the years of field work is an imperative, not only needed for the folklore scholars, but also for broad promotion of this texts. The Indexes of tales plots play the most important role in this structuring – in fact the thesauri of the national ideas about the human and the world surrounding. The task of thesauri compiling could not be regarded as fully solved despite the existence of well-known and successfully used Indexes of fairy stories. The need for methods of unsupervised or semi-supervised classification with the participation of folklorists, ontology, taxonomy of raw material is obvious. This article presents approaches and tools, including the method of dynamic programming, multi-dimensional scaling and hierarchical clustering, designed to facilitate the solution of this problem.

Existing Indexes

Folk stories are usually classified according to the system introduced by the Finnish scholar Aarne [Aarne, 1910]. Folklore Indexes for different national sources were made according to that system. Such Index automatically enters into the scientific circulation the raw material, newly collected or recovered from the written sources. Aarne systems attributes each "fairy tale" to a particular plot schema, disregarding particular plot passages and characters - if in a fairy tale tells about a devil, his role is could not be replaced by, say, the sorcerer, although the functions of these characters as a whole will coincide. Aarne’s Index received the broad distribution and is used by the scholars all over the world. It was translated into many European languages also it was translated into Russian by N.P.Andreev [Andreev, 1929]. Andreev, while translating Aarne’s Index into Russian has found that a large number of Russian fairy tales are not presented in the Index prompting the need of its expanding.

In Thompson’s system [Thompson, 1973], the development of Aarne’s system, the complex plots are divided into several elements and for each element different implementations are presented. As a result, the Index is quite malleable, open for extensions and additions. Besides in the Thompson’s Index each plot is supplied with a list of its constituent elements, which provides cross-reference. This book became a universal international directory of fairy stories and each folklorist can’t avoid referring it. Unfortunately, the most representative

Russian tales Index, CYC: [Barag, et al, 1979] does not include Thompson's additions, but this Index is equipped with a significant appendix: a list of the most common stories contaminations.

Development of Indexes

The new tales Indexes are constantly drawn up for different national traditions and genres. Various Indexes of folklore genres in Russian, including electronic exist, the most comprehensive of which is the web resource, http://ruthenia.ru/folklore. It is interesting to consider the index of childish «Horror stories», http://www.unn.ac.ru/folklore/sukaz.htm. The computer system 'SKAZKA' - TALE was created in the Laboratory of Lexicography automation, NIVC, Moscow State University, [Rafaeva, 1998] which was implemented using the Starling database. It provides an opportunity to respond to the various types of requests concerning fairy tales, described in the Index.

The need for clustering

The aforesaid Indices are based on the intuition of the scholars. The formal methods of a plot designation to some index category. We use the cluster analysis methods for solving this problem. Now the article discusses the automatic clustering methods for building Classification of Russian fairy tales plots. We use the material available at http://www.ruthenia.ru/folklore/sus/ hereinafter referred to as "SUS". This resource allows getting more than 4000 html pages with description of the tales and is equipped with references that identify each plot as *andr_87.htm*. The resource was processed for inclusion in a relational database with the main table containing fields: fname – the plot ID and memo field, taleplot - containing the text of the story. A derived table – Taleww obtained from the first one contains the field href – the name of the plot, nfword – the serial number of the word in the plot, tword - the word itself, grame – POS, lemma – word lemma.

The initial data were lemmatized and POS tagged. A complete dictionary of word forms vs. d lemmas for morphological analysis was prepared using STARLING (Http://starling.rinet.ru/downl.php?lan=ru#dict). It should be noted that the lemmatization results not always coincide with the lemmas given in the "word list to the descriptions of plots" – the part of SUS. This is due to the complexity and often inability to perform precise lemmatization disregarding context. In cases of doubt we use a tuple of all possible lemmas for the given word form. Total number of different lemmas in our glossary is 7120.

We tried to identify specific "tale-belonging" terms by comparing the frequency of lemmas in the glossary of fairy stories and the frequency in the of general word-frequency vocabulary (Http://www.artint.ru/projects/frqlist.php). However, these selected terms were not used in the analysis. The distribution of POS tokens is shown in Table 1.

Table 1 Tokens - POS distribution in SUS

POS	Word #	Percent
Noun	28214	37
Adjective	4281	6
Verb	16624	22
Adverb	1723	2
Service words	8152	11
Unknown words	3305	4
Ambiguous POS	13815	18
Total	76114	100

According to the above table we may restrict our analysis to nouns, adjectives, verbs and adverbs, accounting for more than half of all tokens. We can also add words, for which part of speech was not defined, it is likely those are archaisms, regional, rare, words but if they met in two plots, these plots are likely to be proximate.

Table 2. The most frequent words in SUS

Lemma	Word #	Lemma	Word #

Жена (wife)	621	Брат (brother)	213
Человек (man)	462	Деньга, деньги (money)	207
Муж (husband)	410	Хотеть (want)	207
Мужик (peasant)	402	Барин (lord)	203
Волк (wolf)	342	Получать (receive)	202
Поп (priest)	313	Убивать (kill)	197
Царь (king)	310	Разбойник (burglar)	195
Лис, лиса (fox)	295	Хозяин (owner)	191
Девушка (maiden)	284	Ребенок (child)	178
Старик (old man)	148	Женщина (woman)	116
Давать (give)	140	Вор (thief)	116
Работник (worker)	139	Идти (go)	115
Возвращать (return)	135	Убегать (run)	112
Находить (find)	131	Царевич (prince)	111
Дерево (tree)	129	Змей, змей (snake)	111
Лошадь (horse)	128	Друг (friend)	110
Помощь (help)	127	Делать (do)	110
Герой (hero)	127	Птица (bird)	109

This list is used to determine the most affluent cluster of plots, as we can assume that the plot cluster which contains, e.g., word «Барин» ("gentleman") with frequency of 213, should contain substantially more elements than one containing word «Избушка» ("Hut") with frequency of 5, and therefore, is more interesting to analyze.

The bag of words method

The first method used for determining the proximity between two stories was the method of "bag of words." Lemmas coinciding in each pair of plots were counted. Words of length less than 3 were excluded (not to take into account the short abbreviations), numerals and service words. Only nouns (including proper nouns), adjectives, verbs (including derivatives), adverbs, and words without determined POS. After sorting on the number of the coinciding words the top of the sorted list of plots was presented to the experts. The results of expert assess were unsatisfactory: plots of 100 pairs with the highest calculated proximity similarity 3% were incompatible, but in a random sampling of plots the percentage of incompatible was as big as 35%.

The method of sequential matching

As the plot of a folk tale develops sequently, it can be taken as a hypothesis, that in a pair of plots the word sequences also should be the same. To test this hypothesis we developed an algorithm of sequences matching based on dynamic programming method (Viterby algorithm).

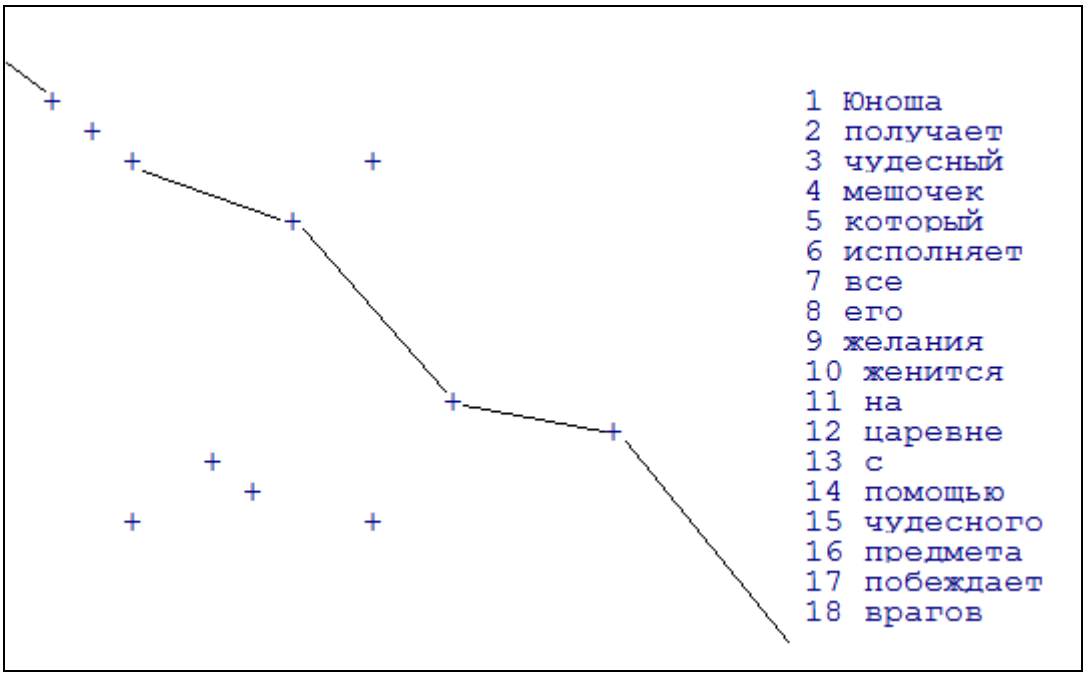

Fig. 1 The sentence on the right is associated with a sentence, «Юноша(1) получает(2) чудесные(3) орешки(4), с(5) помощью(6) которых(7) заслужил(8) чудесных(9) коней(10); на(11) конях(12) допрыгнул(13) до(14) царевны(15) и(16) получил(17) ее(18) руку(19)»

("The young man (1) receives (2) magic (3) nuts (4) with (5) the_help_of (6) which (7) he_deserves (8) wonderful (9) horses (10); on (11) horses (12) he_jumped (13) to (14), princess (15) and (16) got (17) her (18), hand (19) '

Algorithm description

In word sets of plots A and B, Wa = {wa1, .. wan} and Wb = {wb1, ... wbm} we look for pairs of coinciding lemmas: Pairs = {w'i, w'j}. The method of Dynamic programming is illustrated in Fig. 1.

Words Wa are placed by X axis in the sequence in which they are exist in the sentence. Words Wb are placed on Y axis similarly.

Points pk = {w'i, w'j} on the plane we put the pair number k. Now, from the sentence beginning point (0 point) to the sentence ending end point (point n, m) one can trace multiple paths via points pk. It should be noted that for the track sections between the successive pairs pk = {wik, wjk} and pl = {wil, wjl} the condition ik <il, jk <jl, should be held, that is, the path should be a monotonically increasing function. This condition ensures the preservation of the order of words in the pair of plots. Among all possible paths the Viterby algorithm chooses one that includes the largest number of pairs of coinciding words. This number is taken as a measure of the proximity of the two plots.

Expert evaluation of plot pairs from the top of the sorted list shows significant improvement of results compared with the results obtained by the words bag method.

Multidimensional scaling

Visualization of the mutual arrangement of objects (in our case – plots of tales) based on their similarity / difference measure allows visualizing the set of plots close to the target one as well as heuristically identifying clusters of nearby objects. Visualization is achieved with the help of multivariate scaling, applicable in the case, if the sample contains thousands of objects (in our case, the number of plot pairs with at least one coinciding word is more than 600 thousand), and, therefore, the object space is rather multidimensional.

The problem of multidimensional scaling (MDS) is as follows. There is a set of objects X = {x1,. . . , Xt}, for which only some pair-wise distances Dij = ρ (xi, xj) are known. For each object xi ∈ X it is required to build its representation – vector x'i in Euclidean space Rn so that

the Euclidean distance dij between xi and xj approximates the original distance Dij as good as possible.

If dimension of the space n = 2 MDS allows to display the sample as a number of points on the plane. Flat representation usually is distorted, but in general, reflects the basic structural features of multidimensional sampling and its cluster structure. Therefore, two-dimensional scaling is often used as a tool for preliminary data analysis and understanding.

In this case, the distances are known only for some pairs of objects, that is, the distance matrix is very sparse. Consequently, it was not possible to use the algorithm of multidimensional scaling that is used, in particular in the software package Matlab and a special program was developed [Kedrova, Potemkin, 2007], based on an algorithm defined in (Http://forrest.psych.unc.edu/teaching/p208a/mds).

The program produces a two dimensional representation of a particular plot neighborhood displaying links between the story neighbors. A link is displayed, if the two stories have at least one coinciding lemma.

The program allows to set a different number of L = link «levels», i.e., on the 1st level the closest neighbors of this object are displayed, on the second level – the neighbors of neighbors, etc. When L = 2, 3, and more the cluster structure of the plot set is detected. Different mapping of the multidimensional space onto two-dimensional plane, represent different views of the clusters. One can set various cluster centers – the initial plots located at the origin of scaling, move through the list of plots, etc.

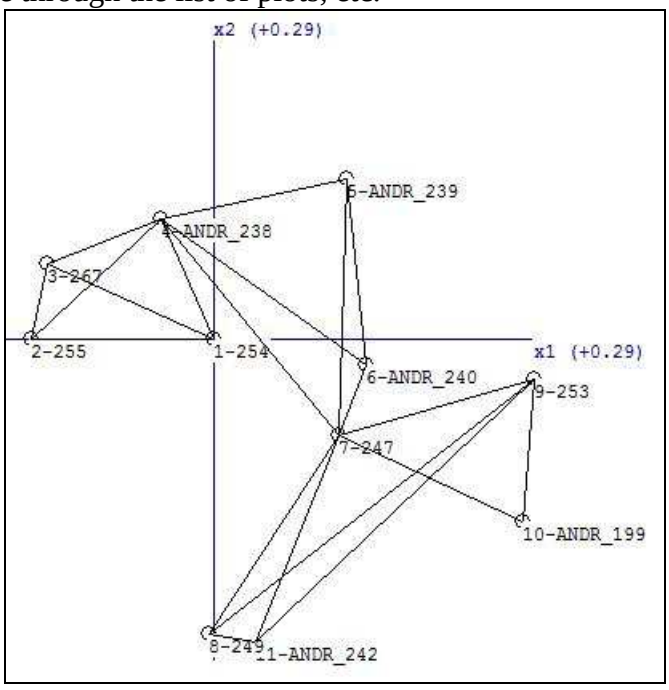

Fig. 2 Neighborhood of story 254: Showing 3 level from point 1-254.

Table 3 Explanation for stories displayed in Fig. 2

1-254	stepmother and stepdaughter: stepdaughter is taken away into the forest; Frost (Baba - Yaga, the devil, in Ukrainian texts often - kobylyachya head) test the girl and award her
2-255	stepmother and stepdaughter: stepdaughter taken away into the forest; Frost (Baba - Yaga, the devil, in Ukrainian texts often - kobylyachya head) test the girl and award her
3-267	stepmother and stepdaughter: stepdaughter taken away into the forest; Frost (Baba - Yaga, the devil, in Ukrainian texts often - kobylyachya head) test the girl and award her
4-ANDR_238	stepdaughter is driven into the forest; Frost (kobylyachya head and so on.) award her for gentleness and politeness, the own daughter is killed.
5-ANDR_239	stepdaughter in the forest; playing hide and seek with a bear, and so on. etc .; A mouse helps her; the native daughter is killed.
6-ANDR_240	stepdaughter in the forest (in the bath, etc...); she makes the forest-devil (devil, etc.) to bring her different things, spending time before the rooster crows; the native daughter is killed (see

	FFS 25, pp. 119 - 120, Sage 31).
7-247	stepmother and stepdaughter: sent by the stepmother to the forest (bath), stepdaughter makes a devil to bring her different thing, spending time before the rooster crows; the native daughter dies.
8-249	stepmother and stepdaughter: stepdaughter is send for the fire to Baba - Yaga; with the help from the wonderful doll she carries out difficult assignments of Baba - Yaga and get the fire; the stepmother and her daughter die.
9-253	stepmother and stepdaughter: stepdaughter drops the spindle into the well (skein into the river) she should go after it; by the way she milking a cow, shaking an apple tree, serves the witch and receives a gift; own daughter also wants to get a gift but doing everything badly and gets a bad gift (she is murdered).
10-ANDR_199	stepdaughter - beautidful, own daughter is ugly (three dwarf, strawberries under the snow, etc).; stepdaughter becomes the wife of the king, she would have a child; stepmother throws queen into the water, and so on.
11-ANDR_242	stepdaughter is sent for the fire to Baba - Yaga; with the help of wonderful doll she performs the difficult assignment of Baba - Yaga and get the fire; stepmother and her daughter die.

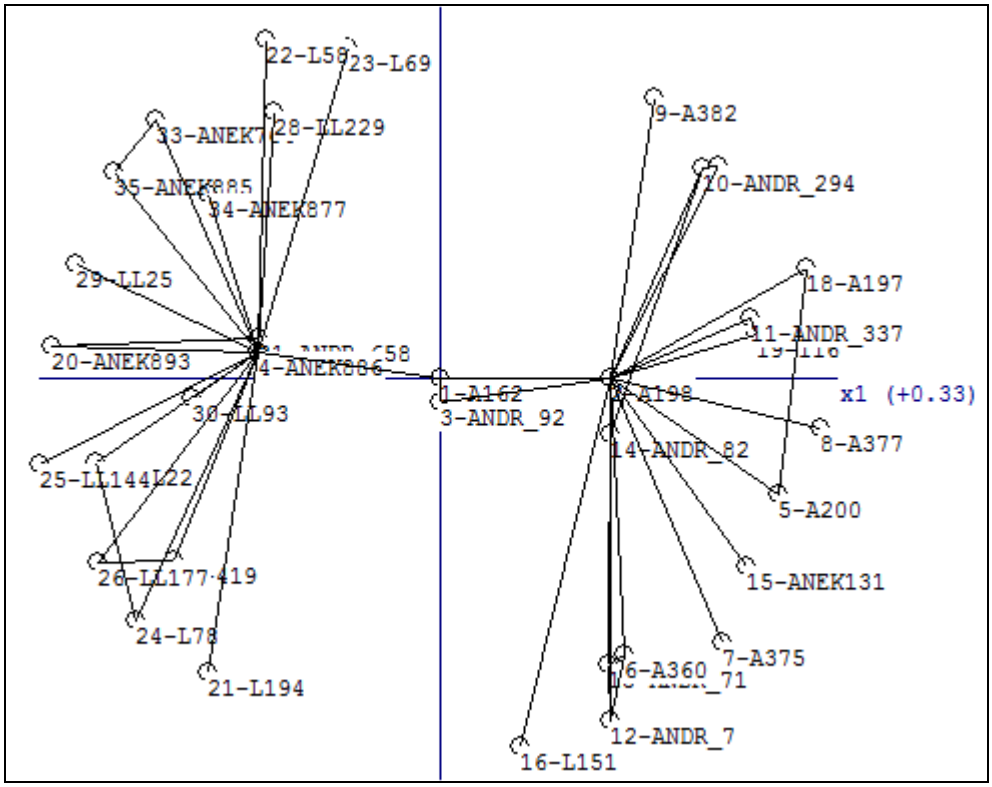

Figure 4. Center: 1 - A161 Faithful Dog unfairly penalized.
Cluster A) 4 - ANEK886 The women worse than devil
Cluster B): 2 - A198 old bread - salts is forgotten

Fig. 4 presents the neighborhood of A161 plot. The neighborhood is obviously divided into 2 clusters, with centers at ANEK886 and A198. Although the link with the center of the original plot cluster is questionable, links within these clusters are mainly traced according to the expert suggestions.

Conclusion

The article describes a new approach to formation of the index of fairy stories based on the analysis of vocabulary, determination the distance between the plots and clustering of plots, by the visual inspection of two-dimensional images obtained by multidimensional scaling. The results can be used for the development of various thesauri.

Bibliography

[Aarne, 1910] - Aarne A. Verzeichnis der Maerchetypen. Helsinki, 1910

[Andreev, 1929] - Andreev N.P., Index of fairy tales plots in the Aarne's system, Leningrad 1929 = [Андреев Н. П. Указатель сказочных сюжетов по системе Аарне., Л. 1929.]

[Barag et. al, 1979] - Barag L.G., Berezovsky I.P., Kabashnikov K.P., Novikov N.V. Comparative Index of subjects: Slavic fairy tale. - Leningrad, 1979 [Бараг Л. Г., Березовский И. П., Кабашников К. П., Новиков Н. В. Сравнительный указатель сюжетов: Восточнославянская сказка. - Л., 1979]

[Kedrova, Potemkin, 2007] Kedrova G.E., Potemkin S.B., Using the corpus of parallel texts to extend specialized bilingual dictionary, // Proceedings of the III International Congress of Russian Language = [Кедрова Г.Е., Потемкин С.Б., Использование корпуса параллельных текстов для пополнения специализированного двуязычного словаря в сборнике *Труды и материалы III Международного Конгресса исследователей русского языка «Русский язык: исторические судьбы и современность»*, с. 627-628]

[Thompson S., 1973] - Thompson S., The Types of the Folktale. Helsinki, 1973

[Rafaeva, 1998] – Rafaeva A.V., Semiautomatic analysis of fairy tales in a computer system SKAZKA // *Proceedings of the International Workshop on Computational Linguistics and its Applications, Dialog'98,* Volume 2, p. 701-706 [Рафаева А.В.,Полуавтоматический анализ волшебных сказок в компьютерной системе СКАЗКА в сборнике *Труды Международного семинара Диалог'98 по компьютерной лингвистике и ее приложениям*,