TRAMES, 2018, **22**(**72**/**67**), 4, 407–424

A LARGE-SCALE STUDY OF WORLD MYTHS

Marc Thuillard[1], Jean-Loïc Le Quellec[2,3], Julien d'Huy[2], Yuri Berezkin[4]

[1]La Colline, 2072 St-Blaise, Switzerland, [2]IMAf UMR 8171, CNRS, F – Paris, France, [3]University of the Witwatersrand, Johannesburg, and [4]Museum of Anthropology and Ethnography (Kunstkamera), Russian Academy of Sciences and European University at Saint Petersburg

Abstract. The study of the narrative elements in tales and myths (motifs) belongs to a long tradition, initially aimed at finding the area of origin of early narratives (Urtexts). This objective, which has been much criticized, is generally abandoned today, but is it possible to establish the basis for an objectively verifiable mythogeography? Computer technology enables sophisticated mathematical computations on databases of an unprecedented scale, which makes it possible to base the comparative mythology on replicable calculation processes. In order to check for several subsets of motifs that could be specific to particular zones or continents, we test here two new methods on a corpus of 2264 motifs from ca. 40.000 myths recorded among 934 peoples around the globe, and we show that these motifs are best classified into two main groups.

Keywords: world mythology, big data, diffusion, classification, phylogenetics, network

DOI: https://doi.org/10.3176/tr.2018.4.05

1. Introduction

We consider here myths as narratives explaining and justifying the present state of the world. They are always regarded as telling the truth in the societies where there are told. The scientific study of this type of story is fairly recent, and it is the subject of a particular discipline: comparative mythology. To facilitate comparisons, the thousands of myths known in all documented societies have been classified into several fundamental types: cosmogonic myths expose the origin of the universe, anthropogonic myths explain the appearance of mankind, ethnogonic myths tell how humanity was divided into different peoples speaking different languages, etc. Many other myths expose the origin of this or that natural or cultural phenomenon: sun, fire, sexuality, domestication, writing, etc. (Le Quellec

and Sergent 2017). Each myth can be deconstructed into 'motifs', defined here as "any features or combinations of features in folklore texts (images, episodes, sequences of episodes) which are subject to replication and found in different traditions" (Berezkin 2015a).

Take, for example, the many myths of the origin of fire (Frazer 1930). It will be easier to compare them if we take into account the presence/absence of motifs such as these: A Living Creature Personifies Fire, Woman Gives Birth to Fire, First Fire is Stolen from Original Owner, Original Owner is a Jaguar, Original Owner is a Toad, etc. The advantage of choosing such motifs as units of analysis is the degree of abstraction it implies. Even if the superficial details of the story have been mistranslated or partially forgotten, the motif is still easy to identify.

First, many mythological motifs remain stable over time and are easily identifyable in similar complex stories at long distances (e.g. Gouhier 1892, Bogoras 1902, Jochelson 1905, Hatt 1949, Lévi-Strauss 2002). Second, the distribution of myths seems to be geographically stable over very long periods of time, as shown for instance by the worldwide contrasting distribution of the 'emergence motif' (i.e. apparition of the first humans from under the earth) and the 'earth diver' motif (Berezkin 2007, Le Quellec 2014). As another example, the myth of the Frog/Toad in the Moon is already documented during the Han dynasty (Dai Lin and Cai Yun-zhang 2005), and propagates over very large distance, from Asia to the northwest coast of North America where it is widespread, without much change. Other motifs are found in similar complex stories and widespread on either side of the Bering Strait (for numerous instances, see Hatt (1949), Berezkin (2013)).

In their studies, folklorists and folk tale specialists generally use Thompson's repertoire of motifs (1955–1958), but this tool is poorly suited to global comparative studies because, for example, Eurasia and North America are over-represented in relation to Africa and Oceania. Thompson has a total of 639 bibliographic sources, and Berezkin no less than 7456, among them more than 2484 are original sources in Russian, mostly about Siberian, Altaic and Finno-Ugric peoples rarely or never mentioned in the motif index. As far as Africa is concerned, Berezkin uses 469 sources, whereas Thompson used only 58.

So, we use here the considerable database of myths elaborated by Berezkin (2015b) which is more comprehensive and better adjusted to mythological studies. This corpus contains over 2264 motifs from over 934 different peoples from all over the world. It was compiled manually and is based on the reading of some 10.000 books, papers and various reports in multiple languages.

A particular myth can be studied in all its details and versions to identify its transformations. This approach allows integrating information from different disciplines, for instance linguistics, anthropology, astronomy, or from ancient written sources. Such work has been done, for example, with the myth of the bird-nester in America (Lévi-Strauss 1964–1971) and Eurasia (Sergent 2009). Alternatively, one may consider a very large corpus of myths or mythological motifs and extract general trends. This last approach has the advantage of facilitating a global

analysis. The difference between the two approaches is equivalent, in the field of genetics, to the difference between the study of a particular gene and a whole genome analysis. A global study of Berezkin's corpus was previously done (Korotayev and Khaltourina 2011, Berezkin 2013, 2017) using Principal Components Analysis (PCA). PCA is a method well adapted to big data but furnishing a limited amount of information in comparison to the methods used in this study. The analysis showed that the different peoples are grouped within clusters corresponding to well-defined geographical regions. It is one of the goals of this work to verify the existence of these clusters with an independent method and to analyse in more details the proximity relationships between the different clusters.

2. Methods

2.1. Phylogenetic approaches

The study of myths using mathematical methods has its roots in their formalization, allowing a structural analysis. The use of biological metaphors (for review, see Hafstein 2001) and of statistics (e.g. Boas 1895:341–347) is very old in comparative mythology. Adler (1987) was the first person to apply phylogenetic tools to classify myths and folktales, followed by Oda (2001) and Tehrani (2013). The phylogenetic method was also used to reconstruct the evolution of myths and traditions (d'Huy 2012, Le Quellec 2015), to study the ecotypification of many variants of a same myth (d'Huy 2013, Ross, Greenhill and Atkinson 2013) and it seems compatible, at least for a part, with the structural approach (Thuillard and Le Quellec 2017). This summary is given for memory, and it is important to note that our own paper moves away from these classical phylogenetic approaches.

After coding, typically with binary characters, the different versions of a myth can be analysed using mathematics or computational methods. The data are coded so that if a motif is present, it takes state '1' while if it is absent it has state '0'. In this sense, each motif can be interpreted as a binary character and each entry (people) as a taxon. The distance matrix between two taxa is computed summing up the distance on each motif. The distance is zero if the two taxa have the same motif's state and one otherwise.

The representation of the different motifs on a phylogenetic tree is based on the following assumptions:

i) Motifs are transmitted unchanged over time and space except for minor transformations that may be compared to mutations. A mutation is defined as the appearance or disappearance of a given motif.
ii) A new motif appears only once.

The condition ii) is a mathematical condition that is seldom perfectly satisfied. A central result in phylogenetic study, applied to myths, is that motifs transforming according to i)–ii) can be exactly represented by a phylogenetic tree (Semple and Steele 2003). Figure 1a shows an illustration of this result. In real applications, if the probability of a motif to appear twice is very low then a phylogenetic tree is often a good representation of the data.

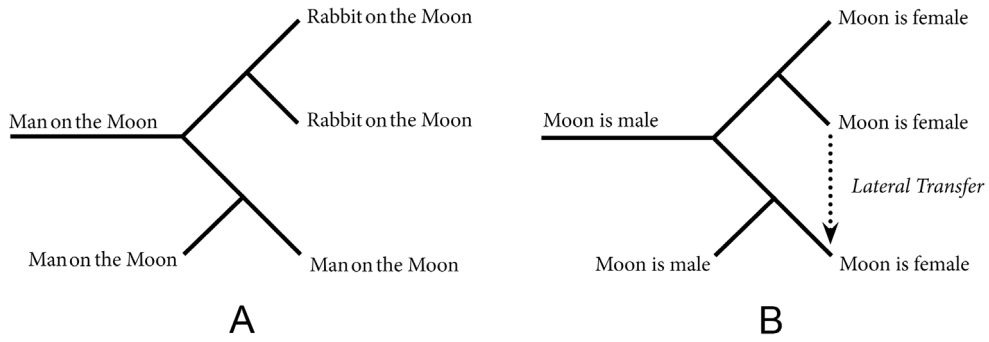

Figure 1. Examples showing the relations between motifs in the case of a) Phylogenetic tree: a new motif appears only once on the shortest path between any two nodes in the tree; b) Phylogenetic network: motifs are transmitted along the branches of the trees but also through lateral transfer (The arrow shows a transfer of a motif).

Unlike genes, cultural elements can be acquired both from other members of the same group of peoples and from outside that group, i.e. they can move from people to people without the need for those peoples to be genetically related. Thus, the distribution of cultural elements and genetic markers will not necessarily co-occur across different populations. Transmission may occur within a population or through cultural interaction between different populations.

In recent years, it has become increasingly clear that a phylogenetic tree is often a too crude representation of the relationships between motifs. A distinct group (i. e. taxon) may inherit motives from several other groups. Figure 1b shows an example in which a motif is inherited both in direct descend as well as through interaction with a distant taxon. This latter process is named in analogy to genetics a lateral transfer. As long as lateral transfers are between adjacent nodes, the different motives can be represented by a phylogenetic network (Thuillard and Moulton 2011). Phylogenetic analysis of data proceeds into two steps.

ii) Order the different taxa. Figure 2 shows, using a simple example, the action of the NeighborNet ordering algorithm.
iii) Validate the data to find out if they fit well to a phylogenetic tree and network.

Validation is done using a contradiction index (Thuillard 2007, Thuillard and Fraix-Burnet 2011). The contradiction index computes a measure of the deviation of the ordered data to a perfect phylogenetic network. Contrarily to global indices, the contradiction can be computed on each separate taxon. The main question behind any comparative study is how to validate the results. In many studies, results are supported by specifics indices showing that on average the results can be trusted. Having phylogenetic studies in mind, a good index does not indicate that the classification is correct in all its details. There is a need for better indices. In this context, the contradiction index is a useful measure that provides both a global and a local indication of the quality of a fit to a phylogenetic description. In this study the average contradiction was moderate, but quite high in Eurasia. For

that reason, no phylogenetic network is shown in this study. We believe that representations of the data as in Figure 4–6 permit to better grasp the underlying structure of the data.

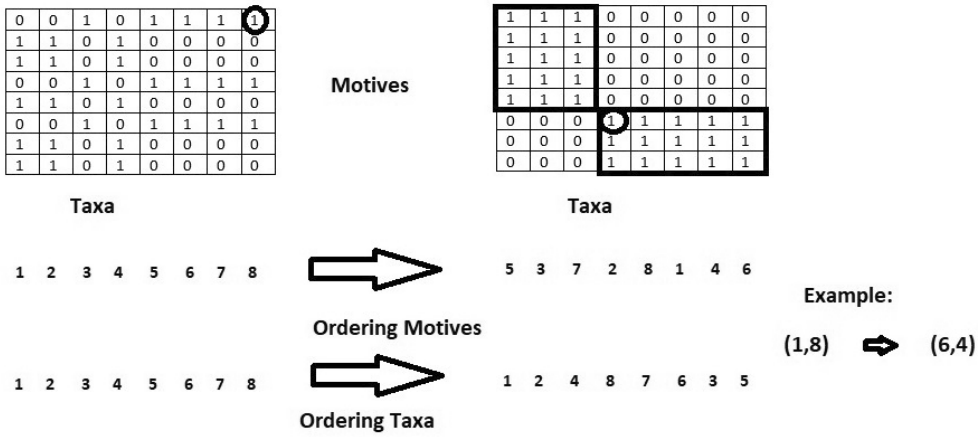

Figure 2. Simple example showing how the NeighborNet ordering algorithm permits to reorder the taxa and motifs so as the data form two clusters. The order is circular in the sense that the first taxon is defined as being adjacent (and consecutive) to the last one.

Due to the very large size of Berezkin's database, standard software programs could not be used in this study and a computationally very efficient variant of the NeighborNet ordering algorithm (Bryant and Moulton 2003) was implemented using the approach in Thuillard and Fraix-Burnet (2009). Contrarily to previous studies (d'Huy 2013, Ross et al. 2013, Thuillard, Le Quellec, d'Huy 2018), NeighborNet was applied to both the taxa (peoples) and the motifs. Anticipating the results, after ordering, one observes that motifs with state '1' have a high density within well-defined clusters. Within a cluster the distribution of state '1' seems to be mostly random. In order to better define the clusters, the data were processed with a correlation operator. This approach is a valid approach on the observed distribution showing well-defined clusters of points after ordering both the motifs and taxa. The correlation matrix $T_{ij} = cor(X_i, s_j)$ was computed with X_i representing after ordering the i^{th} taxon and $s_{ij} = cor(Y_i, Y_j)$ the correlation between each pair of characters (Y_i, Y_j). The different clusters were then identified by a segmentation algorithm using a Laplacian operator (Al-Amri, Kalyankar and Khamitkar 2010). All clusters are observed in a large group (665 taxa and 1477 motifs) of adjacent taxa and motifs characterized by a low average contradiction value on the ordering of both taxa and motifs. In order to analyse how clusters relate with each other, the distance matrix was averaged over each cluster and represented in a gray-scale heatmap (Figure 4). The remaining 787 motifs were analysed in a second classification using all taxa. In order to compare both corpora

corresponding to the two classifications, the frequencies of the words composing the motifs (n = 175.092) were analysed by summing up the number of occurrences of a state '1' for each character used in the first (resp. second) classification (Table 2).

2.2. Area study

In the present case, we were confronted with the difficulty that well-defined clusters are identified but the relationships between these clusters are difficult to establish as the distribution of the different states connecting the clusters does not always fit well to a phylogenetic network (see Fig. 4 and related discussion on the contradiction index). For that reason, a different approach was developed. The method uses directed graphs to represent proximity relations between clusters. The use of digraphs as an extension of phylogenetic networks is known (Bordewich and Semple 2007) but the application to the study of myths is new.

A matrix M with as many lines as clusters and as many columns as characters was built by averaging over each cluster and character the number of taxa with state '1'. A digraph (i.e.: directed graph) was generated by constructing a proximity matrix P_{ij} between pairs of clusters. The proximity matrix was first initialized to zero. A recurrent formula was used on each character to compute the proximity matrix. For each character, one has

$$P_{ij}(k+1) = P_{ij}(k) + 1. \quad (1)$$

If the cluster (i) has the highest average value on all clusters and the cluster (j) has the second highest value above a given threshold (0.03 in this study), otherwise $P_{ij}(k+1) = P_{ij}(k)$. Figure 3 illustrates the algorithm.

The higher the weights of a directed edge, the more connected are the two clusters. A large imbalance between the weight of the two directed edges connecting the same two nodes indicates that shared motifs are much more frequent in one of the clusters than in the other one. If more than two clusters fulfil the condition for updating the above proximity matrix, then the two characters are simply ignored and no update is done. Using that supplementary uniqueness condition, one can show that, given a perfect phylogenetic network and after having partitioned all taxa into subset of consecutive taxa, the edges of the digraph are only between taxa that are adjacent on a circular order. This follows directly from the result that binary data can be exactly represented by a phylogenetic network provided the taxa with state '1' are consecutive (Bandelt and Dress 1992). In the result section, we will see that this condition is not fulfilled and that a phylogenetic network is here not the proper representation of the complex structure of the data (Figure 4 is a therefore a better representation than a phylogenetic network as discussed below).

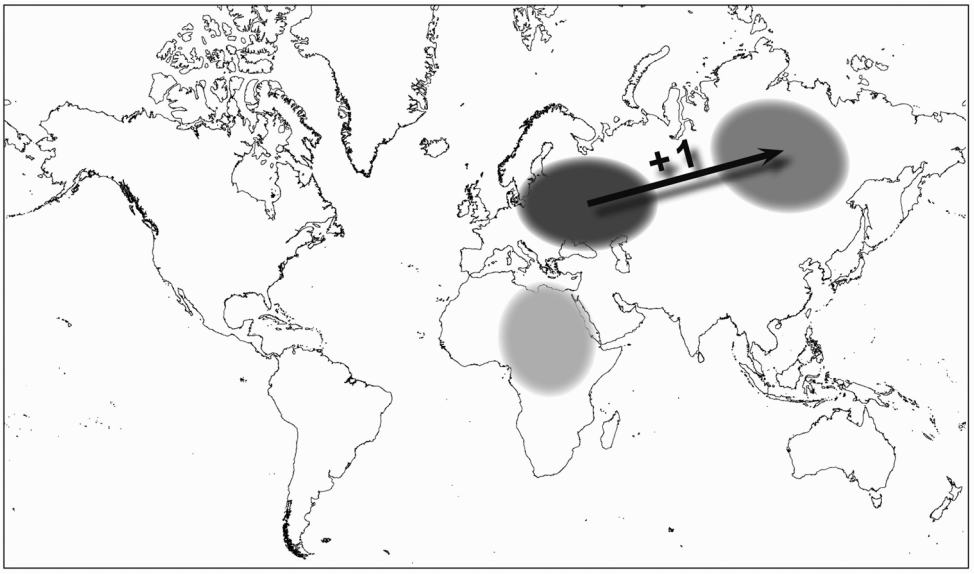

Figure 3. Illustration of the algorithm on one character. The level of grey indicates the percentage of taxa with state '1' on a given character. The arrow relates the cluster with the highest percentage to the cluster with the second highest percentage of state '1'. The proximity matrix is updated accordingly.

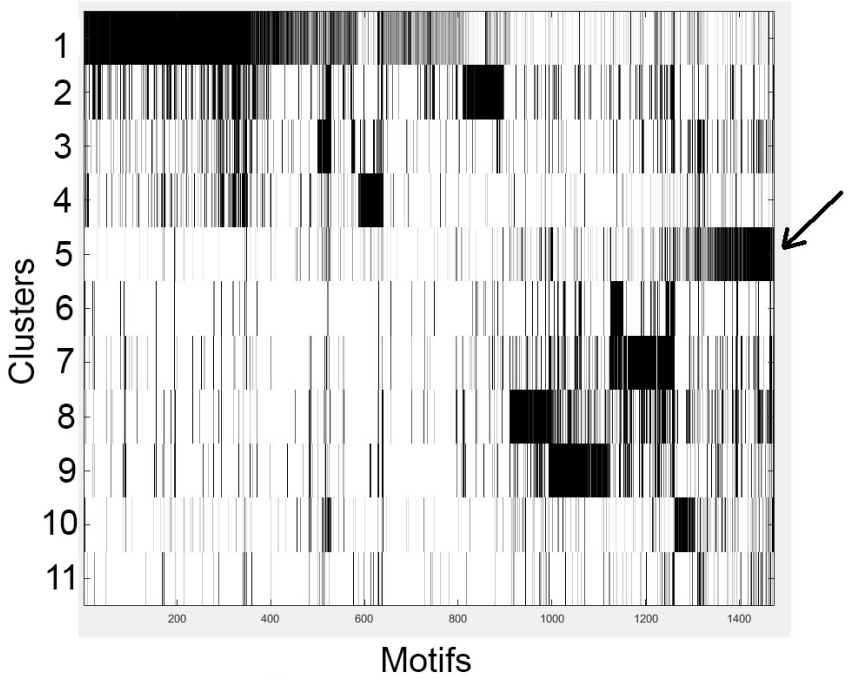

Figure 4. Graphic representation of the average value on each of the 11 clusters (White: zero; Black: value larger than 0.15). As an example, the arrow points to frequent motifs in cluster 5.

3. Results and discussion

3.1. Cluster analysis

The ordering procedure was applied to both the taxa and the motifs' space. Figure 4 shows the value of 1477 consecutive motifs averaged on each of 11 regions. The remaining motifs and taxa did not show any clear structure (i.e. cluster) in this first analysis. The 11 clusters consist each of consecutive taxa after ordering with NeighborNet.

Each cluster in Figure 4 is characterized as well as possible below:

(1) Eurasia, North- and East Africa
(2) Circumpolar Eurasia
(3) Southeast Asia (part of Oceania)
(4) Sub-Saharan Africa
(5) South America (Papua, New Guinea)
(6) Circumpolar America
(7) Bering Strait
(8) Northwest N. America
(9) Central-and East N. America
(10) Pacific Coast (South- and North America), Mesoamerica
(11) Oceania

The different clusters identified in this study correlate very well to the ones obtained with previous studies on the same data (Berezkin 2017).

Eurasia: The main geographic division in Eurasia is between circumpolar regions, Southeast Asia and the rest of Eurasia.

Africa: The continent is divided into two regions: North Africa is within the Eurasian cluster (1), while Sub-Saharan Africa forms a specific cluster. The 'Sub-Saharan Africa' cluster (4) shares a number of motifs with Eurasia.

Oceania: The 'Oceania' cluster contains a grouping of taxa from Oceania and the Pacific Islands. This cluster includes remote islands like Tahiti or Hawaii that were first inhabited recently when compared to other parts of the world.

America: The cluster 'Circumpolar America' contains mainly peoples from circumpolar regions in North America and also in Eurasia around the Bering Strait and Greenland (Eskimo, Netsilik, Iglulik, Caribou, Reindeer and Maritime Koryak). The most common word in this cluster is the word 'raven' who is one of the main character of this mythology. The 'Northwest N. American' cluster corresponds to peoples from the Northwest (with a majority of Salishan, Penuti and Na-Dene speakers) while cluster (9) corresponds to peoples east of the Rocky Mountains speaking languages from different families (Algic, Caddoan, Sioux-Katawba). The 'Pacific Coast' cluster is composed of peoples whose language belongs mainly to Quechua, Uto-Aztecan, Mixo-Zoquean, Oto-Mangean, Mayan. Upon further examination ones observes two sub-clusters: the first one corresponds to Meso-and North American peoples while the second one is a mixture of peoples from Peru, Ecuador and Central America. The Pacific Coast cluster contains also a number of taxa from India and Southeast Asia.

The ‘South America (Papua, New Guinea)’ cluster (5) contains several taxa among Papuans, Solomon Islanders, and South East Asian hunter-gatherers. This suggests a hypothetical Papua / New Guinea / South American ‘supercluster’ already discussed by several anthropologists, such as Nichols (1994) for language, and Gregor and Tuzin (2001) for genders. An over-proportional number of motifs in this cluster are related to body parts ('Body anomalies of the first people’, ‘Body anomalies of inhabitants of a distant land’, ‘No-anus people’, etc.), in particular genitals, as well as to 'woman'. Is it the result of an early or a recent migration or a convergence due to similar habitats? (see e.g. Malaspinas et al. 2014, Raghavan et al. 2015). We will not tackle here the problem of common origin and diffusionism vs convergence as these topics have been treated in much depth and with great insight in Gregor and Tuzin’ s edited book: ‘Gender in Amazonia and Melanesia’ (2001).

Clusters are of different sizes both geographically and in terms of the number of motifs. The ‘Eurasia, North and East Africa’ (1) cluster covers all Eurasia with the exception of circumpolar Eurasia and Southeast Asia. No clear fine structures are observed within this cluster. There is a plausible explanation for that result. Within most of Eurasia, a large proportion of motifs may have diffused quite randomly. A quite different situation is observed in North America which is divided into several small regions with motifs specifics to each region.

Figure 4 is the basis for a more in-depth study of the relationships between clusters. Without being too technical, a basic property of phylogenetic tree or network is that, for a given motif, the taxa with state ‘1’ should be adjacent leading to a zero contribution to the contradiction index. One computes from Figure 4 that the contradiction index is quite large on the ‘Eurasia, North- and East Africa’ cluster (about 20%) and low (on average below 12%) for the North American clusters (6–9). The North American clusters can be well described in first approximation by a phylogenetic tree or network, while the Eurasian clusters are quite far from a phylogenetic structure.

Table 1. Contradiction value for each cluster in Fig. 4

Eurasia		SE Asia	Sub-Sah. Africa	South Am.	North America				Pacific Coast
Circum Eurasia									
1	2	3	4	5	6	7	8	9	10
0.20	0.17	0.13	0.12	0.14	0.12	0.11	0.11	0.13	0.13

Lévi-Strauss (1964–1971) emphasizes at different moments in his career that myths can be related through a complex set of transformations summarized in the so-called canonical formula. For narratives (that are not myths), Mosko (1991) claims that another formula should be used instead. A new perspective on this question has been recently formulated (Thuillard and Le Quellec 2017). Both the

canonical and Mosko's formula have a simple interpretation within the graph theory. The canonical formula describes an instance of myth's evolution that can be described exactly by a perfect phylogenetic tree. Mosko's formula describes a completely different scheme of evolution. It is typically the result of a fast evolution of mythemes resulting possibly in all combinations of binary characters, and Mosko's formula leads to a highly connected graph. Our results show that depending on the regions and the scale at which the data are considered, the best description of the data may change quite drastically. At the scale of a cluster, the motifs are, on average, randomly distributed. We have found that in some regions, for instance North America, motives can be described to a good approximation by a phylogenetic tree or an outer planar network, a type of phylogenetic network (Bryant and Moulton). This topology suggests that motives have diffused among several regions without much transformations and that identical motives are not the result of convergence processes or multiple independent creations. In other areas like Eurasia, motives seem to be randomly distributed among the many peoples as expected from Mosko's formula for narratives. The network describing the distribution of motives is highly connected. Such a network is characteristic of peoples having interacted extensively over eons. In other words, our results show that Mosko's formula for narratives applies, in many instances, to the description of myths.

3.2. Connections between the clusters

As a phylogenetic network cannot be used to represent the whole dataset, one understands the need for a proximity analysis to represent the relationships between clusters. Figure 5 shows the result of the area study. One observes two super-clusters characterized by large weights (broad lines in Figure 5). The first supercluster contains Africa and Eurasia. The second supercluster contains the North American taxa. The two superclusters are connected through the circum-polar Eurasian cluster. The different clusters are also consistent with results from previous phylogenetical (e.g. d'Huy and Berezkin, 2017), statistical (e.g. Bogoras 1902, Korotayev and Khaltourina 2011) and areal (e.g. Berezkin 2013, Le Quellec 2014) approaches on much smaller corpora.

The main information contained in Figure 5 is summarized in Figure 6 (Top).

Repeating the classification on the remaining group of characters results in a second classification represented in Figure 6 (bottom). Geographically, similar groupings are observed but somewhat blurred. For instance, the different North American regions cannot be well differentiated. The number of edges for the sub-Saharan Africa and Oceania clusters were below the threshold of the value P (See method section. No edge is represented if $P<9$). Also, no taxa in Africa could be validated, and only one in Australia. The second group of motifs, associated to the second classification, contains about 200 motifs which contrarily to the first group of motifs are widely spread over several clusters. It is as if superposed to a body of motifs concentrated mostly on a well-defined cluster, a second group of motifs was broadly shared by many peoples.

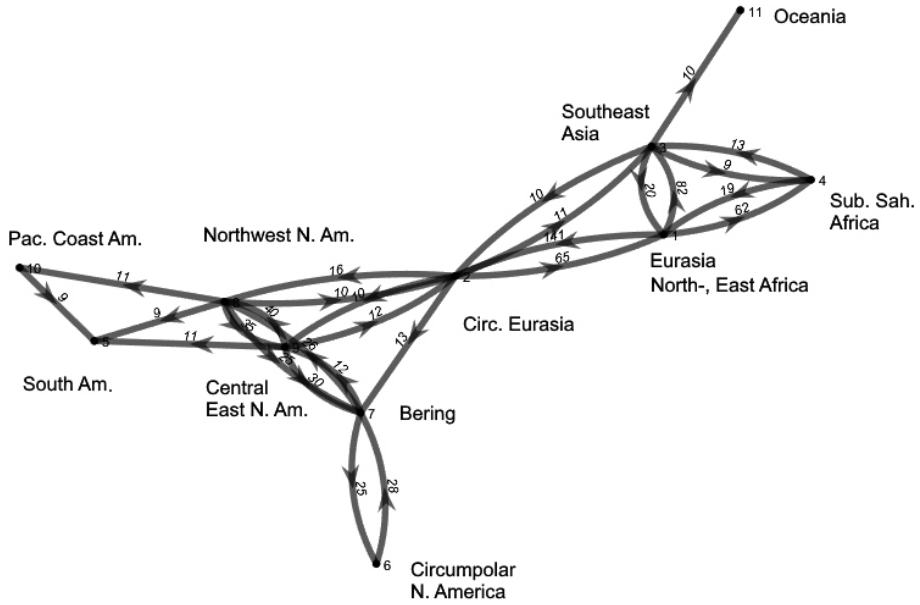

Figure 5. Result of a proximity analysis between clusters in term of density of characters. The weight on the directed edge shows the number of motifs with the two largest frequencies. The arrows point toward the clusters with the second largest frequency. The width of the line is related to the weight on the edge.

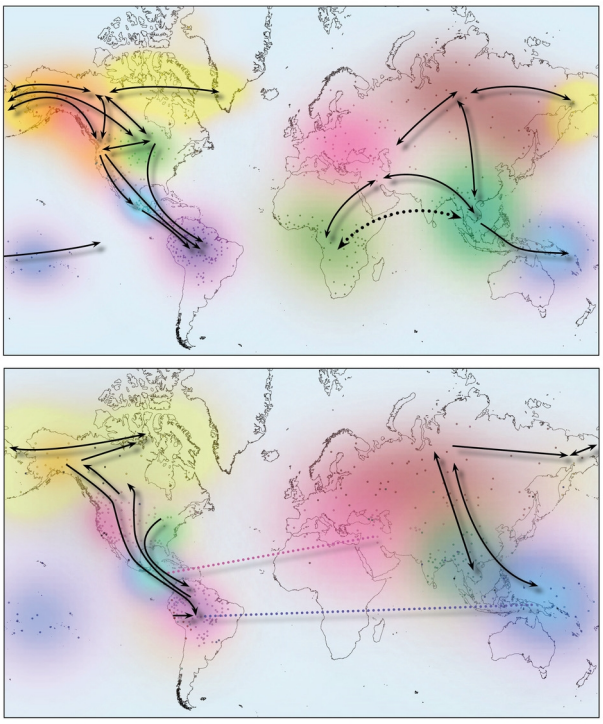

Figure 6. Top: classification of the different taxa in different clusters (one colour per cluster) together with information from the areal study in Fig. 2. Bottom: result of the classification on the remaining motifs. The figure was done using Cartographica 1.4.8.

3.3. Comparison of the two classifications

Figure 7A shows that the most frequent words in the first classification (in order of decreasing frequency: woman, man, animal, people, snake, sky, person, bird, wife, girl) are quite different from the most frequent ones in the second classification (moon, sun, earth, water, animal, man, female, trickster, bird, eclipse).

The first group contains many motifs related to creation myths, the origin of death and a number of well-known motifs such as the rainbow snake and the cultural hero. The second part of the corpus contains several motifs connected to the sun and the moon, to celestial bodies or to a flight in the sky (cosmic hunt, man in the moon, obstacle flights, extra suns and moons annihilated) as well as to the trickster theme. Each subset has two prevalent motifs: Woman+Animal vs Moon+Sun, and they also differ in their level of thematic dispersion.

In both classifications, taxa in South America are related to taxa in Melanesia. The connection is the strongest in the first classification with the South American cluster including 9 Melanesian taxa. Many motifs in the South American cluster are related to 'body parts' and to 'woman', two common topics in the first classification. Fig. 7B shows that in the second group the highest frequencies concern only twenty themes, while all other motifs (those in green) appear very rarely, or only once. On the other hand, the first group has a lower proportion of rare motifs, the dominant themes being much more numerous than in the other corpus. This demonstrates that the second group is far less "diversified" than the other, and this all the more remarkable considering that most of its motifs have a very extended geographical distribution.

Table 2 shows the 10 most frequent words in each corpus with their number of occurrences.

Table 2. Words with the largest number of occurrences

Corpus 1		**Corpus 2**	
woman	1156	moon	2508
animal	1032	sun	2197
man	966	trickster	949
death	963	man	913
bird	832	water	741
wife	801	male	697
sky	675	earth	690
girl	660	fox	669
snake	639	animal	656
turn	606	female	564

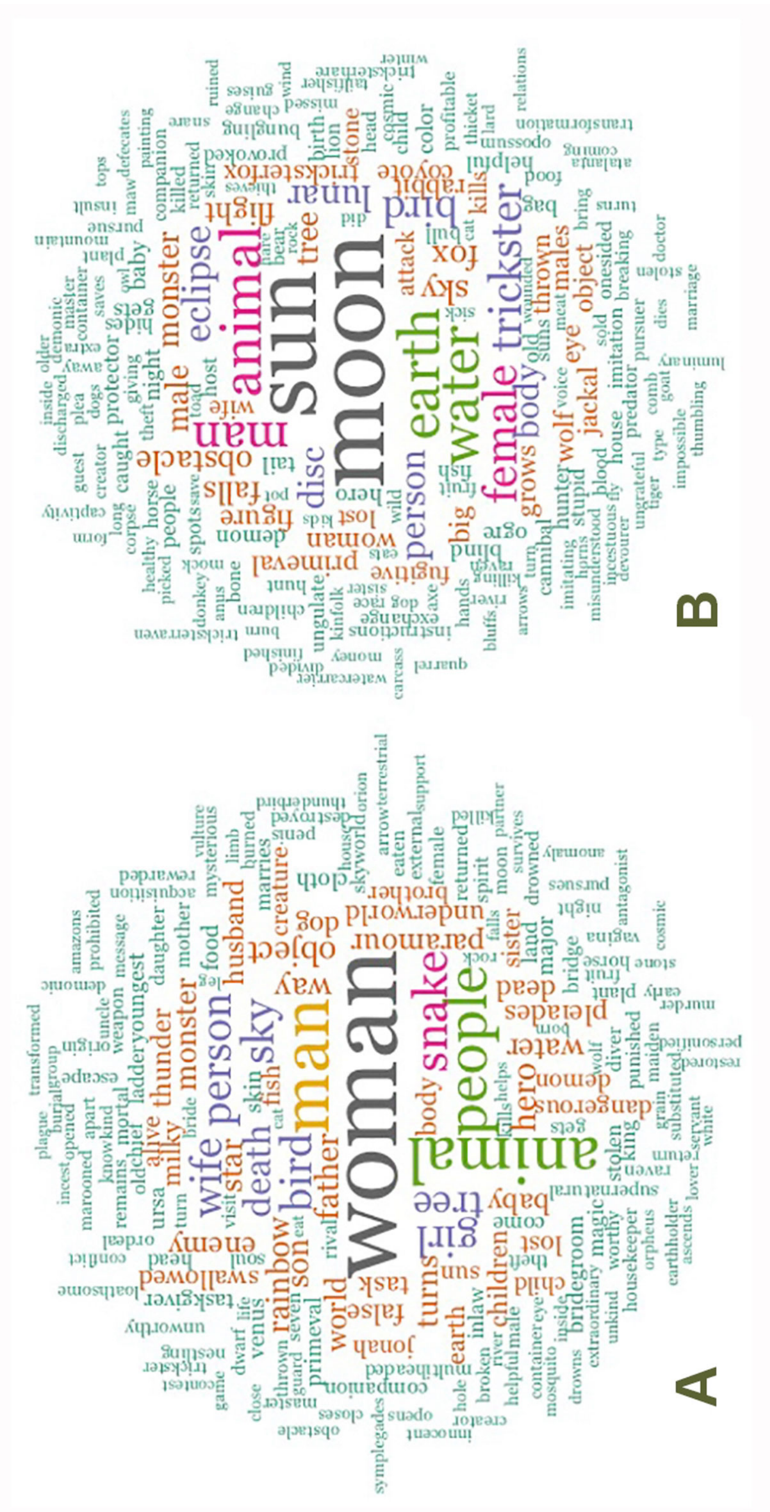

Figure 7. Word frequency for the most frequent words, with the font size proportional to the frequency a) first classification, b) second classification. The word clouds were built with the 'tm', 'SnowballC', 'wordcloud', 'RColorBrewer' and 'pluralize' libraries in R. Stopwords were removed. Plural and singular were treated as equivalent (moon vs moons). We have kept all the items. Stemming and lemmatization makes the dichotomy even more spectacular.

Three motifs are found in most clusters (≥ 8/11). The two first motifs are 'Colours of bird' and 'White raven'. Once the raven is replaced by a local black bird, the motif extends to South America. The white raven motif is a very old one, often part of the Flood and the Earth diver myths, that got transferred to the New World in pre-Columbian times (Korotayev *et al.* 2006). The third motif is 'The Hole in the Sky', which is a concept related to the vision of a solid sky, at least 4000 years old (Seely 1991). More generally, one observes a good correlation between widely shared motifs and myths that are documented in ancient written sources.

4. Conclusions

There are essentially two main corpora of motifs that are geographically intertwined. On both corpora, one observes a very good correspondence between the different clusters obtained after classification and the collection area of the motifs. Let us mention two particular results. On the first corpus, a clear connection is seen between America and Eurasia through the circumpolar regions in the first group of motifs. The association between myths found among peoples in South America and in Papua, New Guinea and the neighbouring islands is not a new observation, but it is difficult to explain. Creation myths, origin motifs and a number of well-known motifs such as the cultural hero or the rainbow snake are a core concern in the first corpus of motifs, while celestial bodies are a central focus in the second one, as well as animals and trickster stories. The most frequent motifs in the second corpus are quite different from the ones in the first corpus, the moon and the sun being a central focus, as well as animals together with the trickster theme. The second corpus contains the majority of motifs having a very extended geographical distribution and includes a large number of motifs present on most regions. Quite interestingly, one observes that the motifs with the broader distribution are often quite old myths that did propagate mostly orally but were recorded in writing at least in one location during ancient time. This confirms the large diffusion and stability of some ancient myths.

In a wider perspective, we have shown that the use of a very large database should enable us to renew and improve the study of the worldwide distribution of mythical motifs and "to test a richer array of hypotheses" (Henrich et al., 2010: 81). This makes it possible, in particular, to avoid the sampling bias observed in previous comparative studies (Bortolini et al. 2017, d'Huy et al. 2017). Our results show that world mythologies are structured in geographical patterns and confirm the existence of great dichotomies like the one between 'Gondwanian' and 'Laurasian' myths in Witzel's terminology (2001, 2012), but they also indicate that the global distribution of myths cannot be reduced to such simple oppositions. Now, it would be very interesting to be able to cross our results with other types of data, for example regarding the environment (e.g. Botero et al. 2014) or the rituals (Gray and Watts 2017), etc., in continuation of comparable research (e.g. Currie

2013, Jordan and Huber 2013, Kirby et al. 2016). The problem is that we are currently facing the non-interoperability of several very large databases built independently of one another, but this type of difficulty should be overcome in the future.

The methods used in this study extend phylogenetic approaches to more complex topologies. Our approach builds a bridge between phylogenetic studies and network analysis (Kenna and MacCarron 2016).

Obviously, the methods presented here are not limited to myths. The differences and similarities between the evolution of genes, languages and cultures have been thoroughly studied (Ross, Greenhill and Atkinson 2013). The conclusions are that despite the important differences between genes, languages and cultural traits, similar theories and methods can be applied to all of them separately. Considering that phylogenetic networks find their origin in the work of the archaeologist Flinders Petrie (1899) it is quite clear that the methods presented may also be relevant to archaeology (Le Quellec 2017).

Address:
Jean-Loïc Le Quellec
Institut des Mondes africains, UMR 8171 CNRS
Brenessard
85540 - St-Benoist-sur-Mer
France
E-mail: JLLQ@rupestre.on-rev.com

References

Abler, Thomas (1987) "Dendrogram and celestial tree: numerical taxonomy and variants of the Iroquoian creation myth". *The Canadian Journal of Native Studies* 7, 2, 1987, 95–221.

Bandelt, Hans-Jürgen and Andreas Dress (1992) "Split decomposition: a new and useful approach to phylogenetic analysis of distance data". *Molecular Phylogenetics and Evolution* 1, 242–52.

Berezkin, Yuri (2007) "'Earth-Diver' and 'Emergence from under the Earth': cosmological tales as an evidence in favor of the heterogenic origins of American Indians". *Archaeology, Ethnology and Anthropology of Eurasia* 4, 32, 110–123. https://doi.org/10.1134/S156301100704010X

Berezkin, Yuri (2013) *Afrika, Migracii, mifologija. Arealyrasprostranenija fol'klornyx motivov v istoričeskoj perspective.* [Africa, migration, mythology. Distribution of folklore motifs areas from a historical perspective.] Saint-Petersburg: Nauka.

Berezkin, Yuri. (2015a) "Spread of folklore motifs as a proxy for information exchange: contact zones and borderlines in Eurasia". *Trames* 19, 1, 3–13. https://doi.org/10.3176/tr.2015.1.01

Berezkin, Yuri (2015b) "Folklore and mythology catalogue: its lay-out and potential for research". In Frog and Karina Lukin, eds. *The Retrospect Methods Network Newsletter* 10, 56–70. Between Text and Practice. Mythology, Religion and Research. A special issue of RMN Newsletter, Helsinki: University of Helsinki.

Berezkin, Yuri (2017) "Peopling of the New World from data on distributions of folklore motifs". In: R. Kenna, M. MacCarron, and P. MacCarron, eds. *Maths meets myths: quantitative approaches to ancient narratives*, 71–89. (Understanding Complex Systems.) Cham: Springer. https://doi.org/10.1007/978-3-319-39445-9_5

Boas, Franz (1895) *Indianische Sagen von der Nord-Pacifischen Küste Amerikas*. Berlin: A. Asher.

Bogoras, Waldemar (1902) "The folklore of Northeastern Asia, as compared with that of Northwestern America". *American Anthropologist* 4, 4, 577–683. https://doi.org/ 10.1525/aa.1902.4.4.02a00020

Bordewich, Magnus, and Charles Semple. (2007) "Computing the minimum number of hybridization events for a consistent evolutionary history". *Discrete Applied Mathematics* 155, 8, 914–928.

Botero, Carlos et al. (2014) "The ecology of religious beliefs". *Proceedings of the National Academy of Sciences* 111, 47, 16784–16789.

Bortolini, Eugenio et al. (2017) "Inferring patterns of folktale diffusion using genomic data". *Proceedings of the National Academy of Sciences* 114, 34, 9140–9145.

Bryant, David and Vincent Moulton (2003) "Neighbor-net: an agglomerative method for the construction for phylogenetic networks". *Molecular Biology and Evolution* 21, 255–65. https://doi.org/10.1093/molbev/msh018

Currie, Thomas (2013) "Cultural evolution branches out: the phylogenetic approach in cross-cultural research". *Cross-Cultural Research* 47, 2, 102–130.

Dai Lin and Cai Yun-zhang (2005) "Hexagram statement 'Gui Mei' written on bamboo slips in Qin dynasty and myth goddess Chang flying to the moon". *Journal of Historical Science* 9,4,

Frazer, James George (1930) *Myths of the origin of fire: an essay*. London: MacMillan and Co.

Gouhier, Charles-Félix-Hyacinthe (1892) *L'Orphée américain*. Caen: Ch. Valin Fils.

Gray, Russel and Joseph Watts (2017) "Cultural macroevolution matters". *Proceedings of the National Academy of Sciences* 114, 30, 7846–7852.

Gregor, Thomas and Donald Tuzin (2001) *Gender in Amazonia and Melanesia: An Exploration of the Comparative Method*. Berkeley: University of California Press. https://doi.org/10.1525/california/9780520228511.001.0001

Hafstein, Valdimar (2001) "Biological metaphors in folklore theory: an essay in the history of ideas". *Arv* 57, 7–32.

Hatt, Gudmund (1949) *Asiatic influences in American folklore*. København: Ejnar Munksgaard.

Heinrich Joseph, Steven Heine, and Ara Norenzayan (2010) "The weirdest people in the world?" *Behavioral and Brain Sciences* 33, 61–135.

Howe, Christopher and Heather Windram (2011) "Phylomemetics – Evolutionary Analysis beyond the gene". *PLoS Biol* 9, 5, e1001069. https://doi.org/10.1371/journal.pbio.1001069

d'Huy Julien (2012) "Un ours dans les étoiles, recherche phylogénétique sur un mythe préhistorique". *Préhistoire du sud-ouest* 20, 1, 91–106.

d'Huy Julien (2013) "A Cosmic Hunt in the Berber sky: a phylogenetic reconstruction of Palaeolithic mythology". *Les Cahiers de l'AARS* 16, 93–106.

d'Huy, Julien et al. (2017) "Studying folktale diffusion needs unbiased dataset". *Proceedings of the National Academy of Sciences Letter.* www.pnas.org/cgi/doi/10.1073/pnas.1714884114.

d'Huy, Julien and Yuri Berezkin (2017) "How did the first humans perceive the starry night? – On the Pleiades". *The Retrospective Methods Network Newsletter* 12–13, 100–122.

Jochelson, W. (1905) *The Koryak: The Jesup North Pacific expedition. Franz Boas, ed*. (Memoire of the American Museum of Natural History, New York.) Leiden: E.J. Brill; New York: G.E. Stechert.

Jordan Fiona and Brad Huber (2013) "Evolutionary approaches to cross-cultural anthropology". *Cross-Cultural Research* 47, 2, 91–101.

Mac Carron, Pádraig and Ralph Kenna (2016) "Maths meets myths: network investigations of ancient narratives". *Journal of Physics: Conference Series* 681, 1, 012002.

Kirby, Kathryn et al. (2016) "D-PLACE: a global database of cultural, linguistic and environmental diversity". *PLOS ONE* 11, 7, e0158391. doi:10.1371/journal.pone.0158391

Korotayev Andrei and Daria Khaltourina (2011) *Mify i geny: Glubokaja istoričeskaja rekonstrukcija.* [Myths and genes: deep historical reconstruction.] Moscow: Librokom/URSS.

Korotayev, Andrei et al. (2006) "Return of the white raven: postdiluvial reconnaissance motif A2234. 1.1 reconsidered". *Journal of American Folklore* 119, 472, 203–235. https://doi.org/10.1353/jaf.2006.0023

Le Quellec, Jean-Loïc (2014) "Une chrono-stratigraphie des mythes de création". *Eurasie* 23, 51–72.

Le Quellec Jean-Loïc (2015) "Peut-on retrouver les mythes préhistoriques? L'exemple des récits anthropogoniques". *Bulletin de l'Académie des Inscriptions et Belles Lettres* 1, 235–260.

Le Quellec, Jean-Loïc and Bernard Sergent (2017) *Dictionnaire critique de mythologie.* Paris: Éditions du CNRS.

Le Quellec, Jean-Loïc (2017) "Phylomémétique et archéologie". *Les Nouvelles de l'Archéologie* 149, 5–14.

Lévi-Strauss, Claude (1964–1971) *Mythologiques* I-IV. Paris: Plon.

Lévi-Strauss, Claude (2002) "De Grées ou de force?". *L'Homme* 163, 7–18.

Malaspinas, Anna-Sapfo et al. (2014) "Two ancient human genomes reveal Polynesian ancestry among the indigenous Botocudos of Brazil". *Current Biology* 24, 21, R1035-R1037. https://doi.org/10.1016/j.cub.2014.09.078

Mosko Mark (1991) "The canonical formula of myth and nonmyth". *American Ethnologist* 18, 126–151.

Nichols, Johanna (1994) "The spread of language around the Pacific rim". *Evolutionary Anthropology: Issues, News, and Reviews* 3, 6, 206–215.

Oda, Jun'ichi (2001) "Description of structure of the folktales: using a multiple alignment program of bioinformatics". *Senri Ethnological Studies* 55: 153–174.

Petrie, Flinders (1899) "Sequences in prehistoric remains". *Journal of the Anthropological Institute* 29, 295–301.

Raghavan Maanasa et al. (2015) "Genomic evidence for the Pleistocene and recent population history of Native Americans". *Science* 349, 6250, aab3884-1-aab3884-10. https://dx.doi.org/10.1126%2Fscience.aab3884

Ross, Robert, Simon Greenhill, and Quentin Atkinson (2013) "Population structure and cultural geography of a folktale in Europe". *Proceedings of the Royal Society of London B: Biological Sciences* 280, 1756, 20123065.

Sergent, Bernard (2009) *Jean de l'Ours, Gargantua et le Dénicheur d'oiseaux*. La Bégude de Mazenc, Arma Artis.

Seely, Paul (1991) "The firmament and the water above". *Westminster Theological Journal* 53, 227–240.

Tehrani, Jamshid (2013) "The phylogeny of little red riding hood". *PLOS ONE* 8, 11, e78871.

Thompson, Stith (1955–1958) *Motif-index of folk-literature: a classification of narrative elements in folktales, ballads, myths, fables, mediaeval romances, exempla, fabliaux, jest-books, and local legends*. Rev. and enl. ed. 6 vols. Bloomington: Indiana University Press.

Thuillard, Marc (2007) "Minimizing contradictions on circular order of phylogenic trees". *Evolutionary Bioinformatics* 3, 237–247. http://journals.sagepub.com/ doi/full/10.4137/EBO.S909

Thuillard, Marc and Didier Fraix-Burnet (2009) "Phylogenetic applications of the Minimum Contradiction approach on continuous characters". *Evolutionary Bioinformatics online* 5, 33–46. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2747132/

Thuillard, Marc and Vincent Moulton (2011) "Identifying and reconstructing lateral transfers from distance matrices by combining the Minimum Contradiction Method and Neighbor-net". *Journal of Bioinformatics and Computational Biology* 9, 04, 453-470. https://doi.org/10.1142/s0219720011005409

Thuillard, Marc and Jean Loïc Le Quellec (2017) "A phylogenetic interpretation of the canonical formula of myths by Lévi-Strauss". *Cultural Anthropology and Ethnosemiotics* 3, 2, 1–12. https://culturalanthropologyandethnosemiotics.wordpress.com/

Thuillard Marc, Jean-Loïc Le Quellec, and Julien d'Huy (2018) "Computational approaches to myths analysis: application to the Cosmic Hunt". *Nouvelle Mythologie Comparée* 4. http://nouvellemythologiecomparee.hautetfort.com/numero-4-no-4-2018/

Witzel, Michael (2001) "Comparison and reconstruction: language and mythology". *Mother Tongue* 6, 45–62.

Witzel, Michael (2012) *The origins of the world's mythologies*. Oxford: Oxford University Press.

Annex “Results of the first classification; First Corpus after classification; Second Corpus after classification; Characters’ list: each taxon is associated to the cluster with the largest density of characters; and List of Edges in Figure 5” is available only in our webpage. DOI: https://doi.org/10.3176/tr.2018.4.06